



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

Volume 109  
Number 8

November 2017

Published eight times

ISSN 0022-0663

# Journal of Educational Psychology

Steve Graham, *Editor*  
Eric Dearing, *Associate Editor*  
Jill Fitzgerald, *Associate Editor*  
Panayiota Kendeou, *Associate Editor*  
Young-Suk Kim, *Associate Editor*  
Beth Kurtz-Costes, *Associate Editor*  
Kristie Newton, *Associate Editor*  
Stephen T. Peverly, *Associate Editor*  
Daniel H. Robinson, *Associate Editor*  
Cary J. Roeth, *Associate Editor*  
Tanya Santangelo, *Associate Editor*  
Malte Schwinger, *Associate Editor*  
Regina Vollmeyer, *Associate Editor*  
Kay Wijekumar, *Associate Editor*  
Li-Fang Zhang, *Associate Editor*

[www.apa.org/pubs/journals/edu](http://www.apa.org/pubs/journals/edu)

CURRENT YR/VOL  
Marygrove College  
McDonough Geschke Library  
8425 West McNichols Road  
Detroit, MI 48221

Editor

Steve Graham, EdD, *Arizona State University*

Associate Editors

Eric Dearing, PhD, *Boston College*  
Jill Fitzgerald, PhD, *University of North Carolina at Chapel Hill*  
Panayiota Kendeou, PhD, *University of Minnesota*  
Young-Suk Kim, EdD, *University of California, Irvine*  
Beth Kurtz-Costes, *University of North Carolina at Chapel Hill*  
Kristie Newton, *Temple University*  
Stephen T. Peverly, PhD, *Columbia University*  
Daniel H. Robinson, PhD, *Colorado State University*  
Cary J. Roseth, PhD, *Michigan State University*  
Tanya Santangelo, PhD, *Arcadia University*  
Malte Schwinger, *Philipps-Universität*  
Regina Vollmeyer, *University of Frankfurt*  
Kausalai (Kay) Wijekumar, *Texas A&M University*  
Li-Fang Zhang, *The University of Hong Kong*

Consulting Editors

Olusola O. Adesope, *Washington State University*  
Mary D. Ainley, *University of Melbourne*  
Patricia Alexander, *University of Maryland*  
Rui Alexandre Alves, *Universidade do Porto*  
Eric Anderman, *The Ohio State University*  
David Aparisi, *University of Alicante*  
Particia Ashton, *University of Florida*  
Shannon Audley, *Smith College*  
Courtney N. Baker, *Tulane University*  
Marcia A. Barnes, *University of Texas*  
Roderick W. Barron, *University of Guelph*  
Sarit Barzilai, *University of Haifa*  
Juliette Berg, *American Institutes for Research*  
David A. Bergin, *University of Missouri*  
Matt Bernacki, *University of Nevada, Las Vegas*  
Ryan P. Bowles, *Michigan State University*  
Lee Branum-Martin, *Georgia State University*  
Michelle M. Buehl, *George Mason University*  
Eric Buhs, *University of Nebraska-Lincoln*  
Matthew K. Burns, *University of Missouri*  
Adriana G. Bus, *Universiteit Leiden*  
Kirsten R. Butcher, *University of Utah*  
Andrew Butler, *Washington University in St. Louis*  
Fabrizio Butera, *University of Lausanne*  
Martha Carr, *University of Georgia*  
Clark Chinn, *Rutgers University*  
Eunsoo Cho, *Michigan State University*  
Sun-Joo Cho, *Vanderbilt University*  
Tim Cleary, *Rutgers University*  
Donald Compton, *Vanderbilt University*  
Pierre Cormier, *Université de Moncton*  
Michael D. Coyne, *University of Connecticut*  
Jennifer Cromley, *Temple University*  
Steve Crooks, *Idaho State University*  
Anne E. Cunningham, *University of California, Berkeley*  
Oliver Dickhaeuser, *University of Mannheim*  
Amy Elleman, *Middle Tennessee State University*  
Andrew J. Elliot, *University of Rochester*  
Steve Elliott, *Arizona State University*  
Carol Evans, *University of South Hampton*  
Ralph Ferretti, *University of Delaware*  
Sara J. Finney, *James Madison University*  
Evan Fishman, *Stanford University*  
Brett Foley, *Alpine Testing Solutions*  
Barbara Fooman, *Florida State University*  
Lynn S. Fuchs, *Vanderbilt University*  
David W. Galbraith, *University of Southampton*  
Colleen M. Ganley, *Florida State University*  
Elizabeth Gee, *Arizona State University*  
George Georgiou, *University of Alberta*  
Amanda Goodwin, *Vanderbilt University*  
Michele Gregoire Gill, *University of Central Florida*  
Art Graesser, *University of Memphis*  
Deleon Gray, *North Carolina State University*  
Barbara A. Greene, *University of Oklahoma*  
Jeffrey A. Greene, *University of North Carolina, Chapel Hill*  
John T. Guthrie, *University of Maryland*  
Antonio P. Gutierrez de Blume, *Georgia Southern University*  
Karen Harris, *Arizona State University*  
John Hattie, *University of Melbourne*  
Michael Hebert, *University of Nebraska—Lincoln*  
Marco G. P. Hessels, *University of Geneva*  
Paul R. Hernandez, *College of Education and Human Services*  
Flaviu Hodis, *Victoria University of Wellington, New Zealand*  
Chris Hulleman, *University of Virginia*  
Mina C. Johnson-Glenberg, *Radboud University Nijmegen*  
Nancy Jordan, *University of Delaware*  
R. Malatesha Joshi, *Texas A&M University*  
Avi Kaplan, *Temple University*  
Carol Anne Kardash, *University of Nevada, Las Vegas*  
Andrew D. Katayama, *United States Air Force Academy*  
Devin Kearns, *University of Connecticut*  
Ben Kelcey, *University of Cincinnati*  
Kenneth Kiewra, *University of Nebraska*  
James S. Kim, *Harvard University*  
John R. Kirby, *Queen's University*  
Noona Kiuru, *University of Jyväskylä, Finland*  
Robert Klassen, *University of York*  
Thilo Kleickmann, *Kiel University*  
Uta Klusmann, *Leibniz Institute for Science and Mathematics Education*  
Terri Kurz, *Arizona State University, Polytechnic*  
Nicole Landi, *Haskins Laboratories*  
Seon-Young Lee, *Seoul National University*  
Pui-Wa Lei, *Pennsylvania State University*  
Hongli Li, *Georgia State University*  
Xiaodong Lin-Siegler, *Columbia University*  
Elizabeth A. Linnenbrink-Garcia, *Michigan State University*  
Min Liu, *University of Hawaii at Manoa*  
Robert Lorch, *University of Kentucky*  
Charles MacArthur, *University of Delaware*  
Joseph P. Magliano, *Northern Illinois University*  
Scott Marley, *Arizona State University*  
Jacob M. Marszalek, *University of Missouri, Kansas City*  
Andrew Martin, *University of New South Wales, Australia*  
Linda Mason, *University of North Carolina, Chapel Hill*  
Lucia Mason, *Università degli Studi di Padova*  
Richard E. Mayer, *University of California, Santa Barbara*  
Matthew T. McCruden, *Victoria University of Wellington*  
Kristen L. McMaster, *University of Minnesota*  
Nicole McNeil, *University of Notre Dame*  
Magdalena Mo Ching Mok, *Hong Kong Institute of Education*  
Paul Morgan, *Pennsylvania State University*

Krista R. Muis, *McGill University*  
P. Karen Murphy, *The Pennsylvania State University*  
Benjamin Nagengast, *Eberhard Karls University of Tübingen*  
John Nietfeld, *North Carolina State University*  
Tim Nokes-Malach, *University of Pittsburgh*  
Nikos Ntoumanis, *Curtin University*  
E. Michael Nussbaum, *University of Nevada, Las Vegas*  
Rollanda E. O'Connor, *University of California, Riverside*  
Yukari Okamoto, *University of California, Santa Barbara*  
Paula Olszewski-Kubilius, *Northwestern University*  
Tenaha O'Reilly, *Educational Testing Service*  
Fred Paas, *Erasmus University*  
Erika Patall, *The University of Texas at Austin*  
Reinhard Pekrun, *University of Munich*  
Harsha N. Perera, *University of Nevada, Las Vegas*  
Yaacov Petscher, *Florida State University*  
Gary Phye, *Iowa State University*  
Pablo Pinaay-Dummer, *Martin-Luther-Universität Halle-Wittenberg, Halle, Germany*  
Isabelle Plante, *Université du Québec à Montréal*  
Jan L. Plass, *New York University*  
Patrick Proctor, *Boston College*  
Karen Rambo-Hernandez, *West Virginia University*  
Katherine Rawson, *Kent State University*  
Lindsey Richland, *University of Chicago*  
Aaron S. Richmond, *Metropolitan State University of Denver*  
Gert Rijlaarsdam, *Universiteit van Amsterdam*  
Bethany Rittle-Johnson, *Vanderbilt University*  
Gregory Roberts, *The University of Texas at Austin*  
Alysia D. Roehrig, *Florida State University*  
Christopher A. Sanchez, *Oregon State University*  
Katharina Scheiter, *University of Tübingen*  
Ulrich Schiefele, *University of Potsdam*  
Dale Schunk, *University of North Carolina, Greensboro*  
Malte Schwinger, *Philipps University*  
Corwin Senko, *State University of New York, New Paltz*  
Timothy Shanahan, *University of Illinois, Chicago*  
Robert Siegler, *Carnegie Mellon University*  
Gale M. Sinatra, *University of Southern California*  
Benjamin G. Solomon, *University of Albany*  
Susan Sonnenschein, *University of Maryland Baltimore County*  
Deborah L. Speece, *Virginia Commonwealth University*  
Birgit Spinath, *Heidelberg University*  
Ricarda Steinmayr, *Technische Universität Dortmund*  
H. Lee Swanson, *University of California, Riverside*  
Keith Thiede, *Boise State University*  
Theresa A. Thorkildsen, *University of Illinois, Chicago*  
Carlo Tomasetto, *University of Bologna*  
Chia-Wen Tsai, *Ming Chuan University*  
Joshua Wilson, *University of Delaware*  
Timothy Urdan, *Santa Clara University*  
Ellen Usher, *University of Kentucky*  
Sharon Vaughn, *The University of Texas at Austin*  
Eduardo Vidal-Abarca, *Universitat de Valencia*  
Candace Walkington, *Southern Methodist University*  
Tanner LeBaron Wallace, *University of Pittsburgh*  
Chris Was, *Kent State University*  
Joanna P. Williams, *Columbia University*  
Christopher Wolters, *The Ohio State University*  
Dana Wood, *Georgia College*  
Friederike Zimmermann, *Kiel University*  
Sharon Zumbrunn, *Virginia Commonwealth University*  
Akane Zusho, *Fordham University*

The main purpose of the *Journal of Educational Psychology*® is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

**Single Issues, Back Issues, and Back Volumes:** For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; call 202-336-5600 or 800-374-2721; or visit [www.apa.org/pubs/journals/subscriptions.aspx](http://www.apa.org/pubs/journals/subscriptions.aspx)

**Manuscripts:** Submit manuscripts electronically through the Manuscript Submissions Portal found at [www.apa.org/pubs/journals/edu](http://www.apa.org/pubs/journals/edu) according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Steve Graham, at [steve.graham@asu.edu](mailto:steve.graham@asu.edu). The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

**Copyright and Permission:** Those who wish to reuse APA-copyrighted material in a non-APA publication must secure from APA written permission to reproduce a journal article in full or journal text of more than 800 cumulative words or more than 3 tables and/or figures. APA normally grants permission contingent on permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$25 per page. Libraries are permitted to photocopy beyond the limits of the U.S. copyright law: (1) post-1977 articles, provided the per-copy fee in the code for this journal (0022-0663/17/\$12.00) is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923; (2) pre-1978 articles, provided that the per-copy fee stated in the Publishers' Fee List is paid through the Copyright Clearance Center. For more information along with a permission form, go to [www.apa.org/about/contact/copyright/index.aspx](http://www.apa.org/about/contact/copyright/index.aspx)

**Disclaimer:** APA and the Editors of *Journal of Educational Psychology* assume no responsibility for statements and opinions advanced by the authors of its articles.

**Electronic Access:** APA members who subscribe to this journal have automatic access to all issues of the journal in the PsycARTICLES® full-text database. See <http://my.apa.org/access.html>.

**Reprints:** Authors may order reprints of their articles from the printer when they receive proofs.

**APA Journal Staff:** Rosemarie Sokol-Chang, PhD, *Publisher, APA Journals*; Mare Meadows, *Managing Director*; Amanda S. Conley, *Journal Production Editor*; Cheryl Johnson, *Peer Review Coordinator*; Jodi Ashcraft, *Director, Advertising Sales and Exhibits*.

**Journal of Educational Psychology**® (ISSN 0022-0663) is published eight times (January, February, April, May, July, August, October, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2018 rates follow: *Nonmember Individual*: \$263 Domestic, \$305 Foreign, \$327 Air Mail. *Institutional*: \$1,020 Domestic, \$1,097 Foreign, \$1,121 Air Mail. *APA Member*: \$127. *APA Student Affiliate*: \$79. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to *Journal of Educational Psychology*, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Effective with the 1986 volume, this journal is printed on acid-free paper.

*Journal of Educational Psychology*® is a registered trademark of the American Psychological Association



---

## Assessment

- 1049 Self-Grading and Peer-Grading for Formative and Summative Assessments in 3rd Through 12th Grade Classrooms: A Meta-Analysis  
*Carmen E. Sanchez, Kayla M. Atkinson, Alison C. Koenka, Hannah Moshontz, and Harris Cooper*
- 1067 Four Semesters Investigating Frequency of Testing, the Testing Effect, and Transfer of Training  
*Donald J. Foss and Joseph W. Pirozzolo*
- 1084 Learning-Related Cognitive Self-Regulation Measures for Prekindergarten Children: A Comparative Evaluation of the Educational Relevance of Selected Measures  
*Mark W. Lipsey, Kimberly Turner Nesbitt, Dale C. Farran, Nianbo Dong, Mary Wagner Fuhs, and Sandra Jo Wilson*

© 2017  
American  
Psychological  
Association

---

## Reading and Math

- 1103 Effects of a Year Long Supplemental Reading Intervention for Students With Reading Difficulties in Fourth Grade  
*Jeanne Wanzek, Yaacov Petscher, Stephanie Al Otaiba, Brenna K. Rivas, Francesca G. Jones, Shawn C. Kent, Christopher Schatschneider, and Paras Mehta*
- 1120 Examining the Relations Between Executive Function, Math, and Literacy During the Transition to Kindergarten: A Multi-Analytic Approach  
*Sara A. Schmitt, G. John Geldhof, David J. Purpura, Robert Duncan, and Megan M. McClelland*

---

## Motivation


- 1141 Achievement Goals, Reasons for Goal Pursuit, and Achievement Goal Complexes as Predictors of Beneficial Outcomes: Is the Influence of Goals Reducible to Reasons?  
*Nicolas Sommet and Andrew J. Elliot*
- 1163 Identifying Pre-High School Students' Science Class Motivation Profiles to Increase Their Science Identification and Persistence  
*Jessica R. Chittum and Brett D. Jones*

Classroom Composition

- 1188 Ethnic Composition and Heterogeneity in the Classroom: Their Measurement and Relationship With Student Outcomes  
*Camilla Rjosk, Dirk Richter, Oliver Lüdtke, and Jacquelynne Sue Eccles*

Other

- iii Acknowledgments  
1204 Call for Nominations  
vi Call for Papers - A Focused Collection of Qualitative Studies in the Psychological Sciences: Reasoning and Participation in Formal and Informal Learning Environments  
1083 E-Mail Notification of Your Latest Issue Online!  
vii Instructions to Authors  
1187 New Editors Appointed, 2019–2024  
ii Subscription Order Information



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

APA JOURNALS®  
Publishing on the Forefront of Psychology

ORDER INFORMATION

Start my 2018 subscription to the  
***Journal of Educational Psychology*®**  
ISSN: 0022-0663

**PRICING**  
APA Member/Affiliate     \$127  
Individual Nonmember     \$263  
Institution     \$1,020

Call **800-374-2721** or **202-336-5600**  
Fax **202-336-5568** | TDD/TTY **202-336-6123**

Subscription orders must be prepaid. Subscriptions are on a calendar year basis. Please allow 4-6 weeks for delivery of the first issue.

Learn more and order online at:  
[www.apa.org/pubs/journals/edu](http://www.apa.org/pubs/journals/edu)

EDUA18



# Self-Grading and Peer-Grading for Formative and Summative Assessments in 3rd Through 12th Grade Classrooms: A Meta-Analysis

Carmen E. Sanchez, Kayla M. Atkinson, Alison C. Koenka, Hannah Moshontz, and Harris Cooper  
Duke University

The “assessment *for* learning” movement in education has increased attention to self-grading and peer-grading practices in primary and secondary schools. This research synthesis examined several questions pertaining to the use of self-grading and peer-grading in conjunction with criterion-referenced testing in 3rd- through 12th-grade-level classrooms. We investigated (a) the effects of students’ participation in grading on subsequent test performance, (b) the difference between grades when assigned by students or teachers, and (c) the correlation between grades assigned by students and teachers. Students who engaged in self-grading performed better ( $g = .34$ ) on subsequent tests than did students who did not. Moderator analyses suggested that the benefits of self-grading were estimated to be greater when the study controlled for group differences through random assignment. Students who engaged in peer-grading performed better on subsequent tests than did students who did not ( $g = .29$ ). On average, students did not grade themselves or peers significantly differently than teachers (self-grades,  $g = .04$ ; peer-grades,  $g = .04$ ) and showed moderate correlation (self-grading,  $r = .67$ ; peer-grading,  $r = .68$ ) with teacher grades. Further, other moderator analyses and examination of studies suggested that self- and peer-grading practices can be implemented to positive effect in primary and secondary schools with the use of rubrics and training for students in a formative assessment environment. However, because of a limited number of studies, these mediating variables need more research to allow more conclusive findings.

**Keywords:** student grading, self-grading, peer-grading, meta-analysis, primary and secondary education

**Supplemental materials:** <http://dx.doi.org/10.1037/edu0000190.supp>

Recent educational reform has emphasized a participatory and collaborative culture of learning in the classroom. Consequently, the popularity of self-grading and peer-grading (SPG) in primary through 12th grade classrooms has increased (Hovardas, Tsivitanidou, & Zacharia, 2014) and, in some instances, become part of school culture (Berger, Rugen, & Woodfin, 2014). As receiving feedback on academic work is an established mechanism through which students learn and achieve (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Butler & Winne, 1995), students themselves can serve as useful sources of feedback via SPG.

SPG involves students making judgments about their own and others’ academic performance. They evaluate the extent to which performance criteria and standards have been met (Boud, 1991) and provide criterion-referenced feedback, that is, grading, to themselves or others. Although SPG in college classrooms has received much attention in the research literature (Atkinson, Sanchez, Koenka, Moshontz, & Cooper, 2016; Falchikov & Boud, 1989; Falchikov & Goldfinch, 2000), less attention has been paid to its implementation in primary and secondary education. The current research synthesis aims to integrate research involving (a) the effects of SPG on subsequent student performance, and (b) the correspondence between student and teacher grades in primary and secondary school classrooms, in particular, differences in average grades given on the same assessment and their distributional similarity (correlation).

Rather than simply measuring student outcomes, educational goals now target evaluative processes that also can improve student performance (Klenowski, 1995). Thus, the “assessment *of* learning” paradigm expanded to include “assessment *for* learning” (Tillema, Leenknicht, & Segers, 2011). In the latter, students are active participants who share responsibility and collaborate with the teacher in the assessment process (Dochy, Segers, & Sluijsmans, 1999). The increased popularity of SPG accompanied this change in focus. SPG embodies assessment for learning in that it requires students to engage in higher level thinking and disciplined inquiry to review, clarify, and correct

---

This article was published Online First March 16, 2017.

Carmen E. Sanchez, Kayla M. Atkinson, Alison C. Koenka, Hannah Moshontz, and Harris Cooper, Department of Psychology and Neuroscience, Duke University.

Carmen E. Sanchez is now at Center for Child and Family Policy, Duke University. Kayla M. Atkinson is now at the School of Social Work, University of Southern Florida. Alison C. Koenka is now at Department of Educational Studies, Ohio State University.

This research was supported by a grant to Harris Cooper by the WT Grant Foundation, titled “The Determinants and Impact of Academic Grades: What Grading Strategies Work Best and Why” (181179). We thank Emily Richardson, Grace Hopkins, and Arielle Kahn for their input on the project.

Correspondence concerning this article should be addressed to Carmen E. Sanchez, Center for Child and Family Policy, Duke University, Duke Box 90545, Durham, NC 27708-0545. E-mail: [carmen.sanchez@duke.edu](mailto:carmen.sanchez@duke.edu)

one's own or others' work. Additionally, as students apply assessment criteria, they develop a clearer conception of the assessed material because of increased exposure to it (Hovardas et al., 2014; Ross, 2006).

### Theoretical Justifications for SPG

The potential advantages of SPG to students can be gleaned from numerous theoretical rationales. These can be grouped according to increases in metacognition, motivation, and transferrable skills (Sadler & Good, 2006; Topping, 1998). Metacognition, put simply, refers to "thinking about thinking," but more broadly refers to the role of executive processes in the overseeing and regulation of cognitive processes (Flavell, 1979). Metacognition can apply to a person's declarative knowledge of one's own learning processes (i.e., metacognitive knowledge) and/or relates to strategies or regulation of cognitive activities (i.e., metacognitive skills). Metacognition has been suggested to be the most powerful predictor of learning and implies that a child has control and knowledge over his or her own thinking and learning activities (Wang, Haertel, & Walberg, 1990). In terms of metacognitive benefits, SPG processes require students to make judgments about their own and others' work, and, as a result, can lead to increased awareness, insight, and reasoning. In particular, self-grading encourages a growth mind-set through an emphasis on revision and progress toward a higher standard of achievement (Andrade & Valtcheva, 2009; Dweck, 1986).

Also, SPG provides students with an opportunity to become directly involved in the assessment process, resulting in a greater sense of autonomy. According to self-determination theory, enhanced autonomy should, in turn, predict heightened intrinsic motivation, or a desire to learn for its own sake (Ryan & Deci, 2000). Classroom assessment theory also predicts that shared ownership of the assessment process will increase effort and achievement through students' increased perceived control and responsibility for learning (Brookhart, 1997).

Additionally, SPG can encourage the development of critical thinking skills about students' own work. Self-monitoring may become internalized and become habit for students (Andrade & Valtcheva, 2009). Finally, involvement in the assessment process could decrease students' cynicism about grading (Evans & Engelberg, 1988) by increasing their confidence that the grade "accurately" reflects learning. When explicit grading criteria are used, expectations for performance become more transparent and students can more clearly understand how a grade was earned, thereby demystifying the grading process (Sadler & Good, 2006).

In regard to transferable skills, SPG may increase communication and collaboration skills, as well as the ability to evaluate future work in professional or academic contexts. Furthermore, the use of SPG can decrease teachers' time spent grading (Sadler & Good, 2006; Topping, 1998).

### Self-Grading and Peer-Grading

Self-grading and peer-grading are discussed together, given that they both pertain to students' involvement in classroom assessment practices. However, some important distinctions exist between the two grading techniques (van Gennip, Segers, & Tillema, 2009). In the continuum of learning between formative assessment

and summative assessment, self-assessment lies closer to formative assessment because it requires the active participation in the judgment of students' *own* work and how it compares with the standard. In essence, self-assessment involves self-regulation and internalization (Andrade & Du, 2007). Furthermore, self-grading assumes a growth mind-set by empowering the students to make corrective changes on their work, thereby underscoring that learning is incremental as opposed to just getting it or not (Dweck, 1986). In the act of self-assessment, students are invited to think about the quality of their own work instead of having someone else evaluate it (Andrade & Valtcheva, 2009).

In comparison, peer-grading requires that a student actively participates in the judgment of another students' work, thereby making peer-grading a fundamentally interpersonal process. The interpersonal nature of peer-grading can be minimized somewhat through anonymous and masked grading (Topping, 2003). Peer-grading provides an opportunity for students to specify the quality of a product of other equal-status students (Topping, 2009) and provides another opportunity to apply what they have learned (Dunning, Heath, & Suls, 2004). Peer-assessment activities can vary across numerous dimensions, in that (a) assessors can be individuals or even pairs or teams of students and (b) the direction of the assessment can be one-way or reciprocal. Importantly, self- and peer-grading are not mutually exclusive and offer potential for triangulation; one can lead to the other and, in turn, inform the other.

Another important consideration when comparing SPG is that self-grading is inherently open to a self-serving bias (Dunning et al., 2004). The "flawed" nature of self-grading relates to students' tendency to overrate themselves because of overconfidence in newly learned skills and students' poor assessment of their own comprehension skills. Peer-grading is not as hindered by students' tendency to be overconfident in their own abilities and provides an opportunity to inform students of shortcomings of which they might have been previously unaware (Dunning et al., 2004).

In addition to whether students are asked to grade their own paper or that of a peer, SPG will be influenced by the student's developmental level as well as teacher training and attitudes. Brief descriptions of these issues and related research can be found in supplemental file A of the online supplemental materials.

### Implementation of SPG

The success of SPG in the classroom may depend on a number of implementation factors. Ross (2006) offered guidelines on how to successfully implement SPG in primary and secondary classrooms: define the criteria (or rubric) by which students assess their work; teach students how to apply the criteria; give students feedback on their grading; and allow students to track their progress to improve performance. Other researchers have added suggestions to Ross's guidelines, such as providing sufficient time for revision after student assessment and using SPG as assessment for learning, rather than assessment of learning (Brown & Harris, 2013). Of note, and perhaps not surprisingly, some have suggested that students are better able to self-grade and peer-grade fact-based tests than tests that require more interpretation and reasoning skills (Bonniol, 1981; Dunning et al., 2004).

To reduce bias and produce clear and meaningful assessment tasks, some researchers have suggested additional environmental



factors to consider when implementing SPG in the K-12 classroom. These include students' awareness of the value of SPG (Goodrich, 1996); a supportive classroom environment that positively influences a child's likelihood to produce and report grading results (Kuncel, Credé, & Thomas, 2005); open discussion between students and teachers; and provision of qualitative feedback (e.g., comments that inform future revisions; Hodgson, 2010). Finally, Tillema et al. (2011) suggested that fairness and transparency should be applied to all steps of the assessment cycle.

Two important components of SPG in the classroom are the use of rubrics and student training to provide structure and guidance. Often, the student's SPG process is marked by incremental steps; first, students are provided a clear rubric with which to grade, followed by examples of how to grade, and then students practice grading. An ideal rubric provides a clear set of criteria and describes varying levels of quality for a specific assignment ("specific" rubric; Andrade & Valtcheva, 2009). Other types of rubrics may provide some criteria reflecting the underlying skills and knowledge within the defined domain, but ultimately leaves the grader to make an overall judgment on the quality of the work ("general" rubric; Lane, 2012). Student training can include modeling, direct instruction, and practice, which are commonly employed classroom practices. For example, teachers may demonstrate rubric use to the class, provide guidance while students engage in the SPG process, and then give students an opportunity to practice SPG independently.

### Previous Reviews of SPG Research

Although several reviews on SPG exist, most focus solely on college or professional school samples (e.g., Topping, 1998) or combine K-12 with college studies in the synthesis (e.g., van Zundert, Sluijsmans, & van Merriënboer, 2010). However, as discussed previously, K-12 and college classrooms should be considered separately because of differences in students' cognitive development and learning environments.

### Meta-Analyses on SPG in College

Despite the importance of focusing separately on K-12 studies, two very similar meta-analyses on SPG in the college setting warrant mention. Falchikov and Boud (1989) aggregated 57 studies that investigated self-assessment compared with instructor assessment. Their results suggested that college students give themselves higher grades than their instructors ( $d = .47$ ) and that student grades demonstrated a moderate relationship with teacher grades ( $r = .39$ ). The quality of the study, course level (i.e., introductory vs. advanced), and subject matter influenced the correspondence between self-grading and teacher-grading. Falchikov and Goldfinch (2000) conducted another meta-analysis with 48 studies on the effects of peer-grading and concluded that students give their peers higher grades than instructors ( $d = .24$ ) and showed a strong intercorrelation among peer-graders ( $r = .69$ ). Design quality, use of rubrics, and the nature of the assessment task (i.e., academic or professional task) moderated the effects of SPG.

### Narrative Reviews on Self-Grading in K-12 Classrooms

Four reviews have narratively examined SPG in the K-12 classroom. Ross (2006) reviewed research evidence on student self-grading, focused largely on the K-12 setting. Evidence on the alignment of self-grading with teacher-grading was weak, with few studies ( $k = 2$ ) directly examining the relationship between self, peer, and teacher grades for a specific outcome measure (that is, mean grades, degree of variation in grades, and/or correlations between student and teacher grades related to the same test). Self-grading was found to generally improve student performance on subsequent assessments (14 of 16 studies reported positive effects).

Brown and Harris (2013) synthesized studies examining the effects of self-grading practices in kindergarten through 12th grade. However, their review is notable in that they broadly defined self-grading by including both studies on more general self-ratings (e.g., van Kraayenoord & Paris, 1997) and "self-rated confidence in accuracy of work" (e.g., Koivula, Hassmén, & Hunt, 2001). The median effect size (ES) fell between .40 and .45 (range =  $-.04$  to  $1.62$ ; no overall ES was reported), with weak to strong correlations between student and teacher distributions of grades ( $r$  range =  $.2$  to  $.8$ ). However, the summary statistics did not distinguish between studies that reported self-grading, self-rating, and self-estimates of performance from those that reported the effect of self-grading on subsequent test scores. Brown and Harris (2013) concluded that self-grading can improve learning outcomes when students are engaged in self-regulation processes (e.g., self-monitoring against objective standards) and when teachers are actively engaged in the development and monitoring of self-grading. Increasing age (and related school experience) appeared to improve the correspondence between students and teachers.

Two meta-analyses have investigated the impact of self-assessment on subsequent writing performance in primary schools. In 2011, the Carnegie Corporation examined the influence of formative writing assessment to improve writing achievement (Graham, Harris, & Hebert, 2011). Based on seven reports, the authors found that when students are taught how to self-grade their own work, scores improved by .46 standard deviations. No moderator analyses were conducted for this study. The authors concluded that self-assessment is an evidence-based practice for improving the writing of American students (Graham et al., 2011).

Another meta-analysis investigated the effects of self-grading of writing assignments on subsequent writing performance (Graham, Hebert, & Harris, 2015). The average weighted ES of self-grading from a total of 11 reports was .62 standard deviations, which indicated a significant impact of self-grading on subsequent performance. A metaregression showed that the quality of the study, feedback structure, or grade level moderated the effect. Both meta-analyses also reported on the effect of peer feedback, but their results were not specific to peer-grading, so studies involving students providing feedback without a grade were also included. Taken together, these reports indicate that the provision of self-feedback, including self-grading, has a positive influence on students' writing achievement.



## Peer-Grading in K-12 Classrooms

Topping (2013) summarized the research on peer-assessment in elementary and secondary schools. He noted that previous literature reviews failed to operationalize peer-grading. A large proportion of studies used survey methodologies, and very few used quasi-experimental designs (two of 16). Notably, the review also pointed to the need for more rigorous methodology in primary studies; no study conducted in elementary school implemented an experimental design, and only one study in secondary schools used a comparison group with a posttest-only design.

Finally, Sebba et al. (2008) conducted a synthesis of research evidence on the impact of students' SPG in secondary schools. Most studies occurred in the United States (62%), and only 11 of 26 (42%) involved comparison groups. Sixty percent of studies (nine of the 15) measuring achievement reported higher scores after use of SPG. Among other suggestions that were consistent with Ross's (2006) recommendations, Sebba et al. (2008) suggested that teachers need instruction in SPG in both initial training and continuing professional development.

Taken together, narrative research syntheses and meta-analyses on both SPG point to the paucity of high-quality empirical studies in K-12 classrooms. Furthermore, most studies focused on the long-term effects of SPG, thereby adhering to the formative assessment view of SPG. Few reviews investigated the degree of grade similarity and/or correspondence between student and teacher grades, suggesting that less attention is paid to SPG as summative assessment in primary and secondary classrooms.

These reviews pointed to some important moderators in SPG. Although self-grading and peer-grading are discussed collectively in this manuscript as "student grading," they are distinct entities that warrant separate examination. Past reviews have pointed to experimental design, classroom characteristics (i.e., students' year in school, course subject), student training, and rubric use to be potentially relevant moderators that might improve SPG outcomes

## The Present Research Synthesis

The present research synthesis fills a void in the SPG literature in relation to the effects of SPG implementation and the correspondence between SPG and teacher grades in primary and secondary classrooms. Our meta-analysis extends previous reviews by updating the evidence base, providing cumulative statistics, and more formally testing for moderators of SPG effects. The primary purpose of this article is to examine whether and when SPG are effective techniques in primary and secondary school classrooms. This includes examining the long-term achievement-related consequences of SPG. It also includes an examination of how correspondent students' grades are to teacher grades with regard to both the mean grades given and their correlation (the placement of a particular student's grade in the grade distribution).

In sum, in order to determine the effects of self- and peer-grading in the classroom, we conducted a series of meta-analyses to examine the following research questions: (a) What are the effects of self-grading and peer-grading practices on subsequent test performance?; (b) What is the mean difference between SPG and teacher grades when grading the same test?; and (c) What is the degree of distributional correspondence between student and teacher graders?<sup>1</sup>

In addition to our main research questions, we also investigated variables that might moderate these relationships. We based our choice of moderators on the claims of scholars who have previously written on SPG and have summarized the research literature. Specifically, we hypothesized that the effect of SPG would be greater

- in secondary than in primary school grades;
- in STEM classes rather than non-STEM classes;
- with students' use of rubrics rather than no definition of criteria by which to score students' work;
- with students' training on how to self- or peer-grade compared with no training;
- with longer training exposures compared with short training exposure; and
- with multiple modes of training rather than either examples or training.

For the first research question examining the long-term implication of SPG, a typical study would have the experimental group complete an assignment, then perform self- or peer-grading to determine a score for the assignment. The control group would not be given any assignment at all or their assignments would be graded by the teacher. After the SPG had occurred in the experimental group (on one or multiple occasions), all students would be given a different posttest that was scored by the teacher or experimenter. The scores from the posttest served as the outcome measure and would be compared between the experimental and control groups to determine the effects of SPG on subsequent tests.

For the second and third research questions examining the relationship between teacher and student grading, the same test would be scored by the teacher and by either the student themselves or a peer. The scores given by the teacher or self/peer would then be compared with each other to determine degree of average grade similarity and the correlation of grades with one another. In these studies, the teacher-graded test and student-graded test would be considered to be yoked outcomes of the control condition and experimental condition, respectively.

With increased emphasis on students' active involvement in the learning process, a clear understanding of SPG in the elementary and secondary schools is needed. We hope to contribute to this understanding.

## Method

### Criteria for Including Studies

A study had to meet several criteria to be included in the research synthesis. First, the study had to focus on one or more of the following research themes: (a) the influence of prior use of students as graders on these students' subsequent test performance, (b) mean score differences when assigned by different graders (i.e., teacher, peer, or self), and/or (c) the correlation between scores assigned by teachers, peers, or self. For the first research theme, reports had to assign some participants to take part in student

<sup>1</sup> Of note, our discussion of student grading purposefully avoided the term "accuracy," as it implies that teacher grades are a perfect reflection of student performance. Instead, we refer to "mean difference," "correspondence," and "correlation" when we make our comparisons.



grading (treatment), whereas others did not participate in student grading (comparison). All students were later tested and graded by a teacher or experimenter. For the second and third research themes, reports had to have both teachers and students grade the same student test and compare either teacher- versus peer-graders or teacher- versus self-graders. All included studies had to either use random assignment of students to conditions or some form of quasi-experimental design. Studies in which students served as their own controls (i.e., pretest-posttest) were not included.

For the purposes of these meta-analyses, we defined the construct of SPG as using criterion-referenced feedback, that is, when students assigned numerical values (or letter grades subsequently converted to numerical values) to either their own or a peer's work in an attempt to make an objective judgment about the quality of the work. In this sense, the grading criterion had to include more than written comment or feedback or general reflection. It had to include a numerical value or grade. Accordingly, SPG requires the students to assess the task directly and systematically use criterion or standards that are task-specific.

We employed three additional exclusion criteria before coding began. First, we excluded studies that did not provide quantitative marking on a specific learning outcome variable, for example, in some excluded studies, students graded their own *general* competence in a subject area (e.g., Ikeguchi, 1996), students provided *perceptions* on how well they or their peers performed (e.g., Wright & Houck 1995), students edited the writing process (and these were tallied) as opposed to providing a measure of a specific learning outcome variable (e.g., Fitzgerald & Markham, 1987; Paquette, 2008), or students *ranked* themselves compared with peers (e.g., Crocker & Cheeseman 1988). Second, we excluded reports that did not have a comparison group (e.g., Andrade, Du, & Wang, 2008). For the first hypothesis, this meant that the study had to have a comparison group that did not partake in self- or peer-grading. For our second and third hypotheses, the same test had to be graded twice, once by the teacher and once by the student—the former served as the comparison. Lastly, we excluded reports that did not provide enough information to calculate an ES (e.g., Beach, 1979; Bickmore, 1981). For reports with unclear methods or missing information to calculate an ES, the study's authors were contacted for additional information. We limited the contact to authors who had recently published (i.e., since 2005;  $n = 3$ ), and two authors responded, which allowed us to include their studies in the synthesis.

## Literature Search Procedures

Our initial database searches sought to identify any studies related to effective grading strategies in general. To do so, we first searched the ERIC and PsycINFO electronic reference databases for published and unpublished documents related to grading strategies. The two databases were chosen because they were most likely to contain reports related to education and developmental differences that might affect instructional practices. The searches were conducted during February 2016 and were not restricted by date of report dissemination. The subject (SU) term “grades (scholastic)” was paired separately in intersection with the following SU terms: “evaluation methods,” “evaluation criteria,” “test methods,” “measurement technique,” “peer evaluation,” “self evaluation,” “multiple choice tests,” “peer grading,” “self grading,” “peer as-

essment,” and “self assessment.” After the initial search, a second search was performed using the same databases and subject terms with the key term “grading (educational)” in intersection with the terms above. Searches were conducted sequentially, with overlapping documents excluded from each subsequent search.

Three coders (two research assistants and a postdoctoral fellow) were trained to examine each report's title, abstract, and keywords from the search results. Each researcher worked independently to categorize each document as to whether (a) it was irrelevant, that is, mentioned grading of tests not at all or only in passing; (b) contained background information on *grading strategies* but was not an empirical study; or (c) included empirical data on the research question of interest.

Within the final grouping of studies, studies conducted using students in Grades K-12 were separated from studies using college students. Studies of *grading strategies* compared different techniques for determining or assigning grades. Typically, they involved the experimental manipulation and comparison of more than one grading strategy. Pretest-posttest designs and case studies of a particular strategy were also included in this category. Excluded documents containing empirical data about grading were those that reported on the grading practices of teachers (without any comparison with student graders), grading systems for program evaluation, or a comparison between online and on-site classes. If at least two coders agreed on a document's placement, it was placed into the agreed-upon category. If the disagreement could not be resolved, the principle investigator was consulted. If relevant reports were misclassified during the initial coding process, two other techniques (e.g., the examination of reference and citation lists in relevant reports, contact with active researchers; see below for more detail) would help to uncover the reports again. In total, 1,459 abstracts were examined and, of these, 323 (22%) were deemed to fit Criteria (b) or (c) within the K-12 domain by at least two researchers.

We then obtained the 323 potentially relevant documents, 94 of which contained empirical data. These reports were examined in their entirety. The relevant reports were further grouped as belonging to specific grading strategies. This categorization suggested that the questions “Does student grading affect later achievement?” and “Do self-assigned and peer-assigned mean grades differ from teacher grading?” had received most of the research attention in the K-12 grading literature. Specifically, we identified 38 empirical studies investigating SPG in the K-12 literature (other studies addressed education-related themes, such as effect of type of test response, effect of grade cutoffs). We subsequently used these 38 studies as the basis to find other relevant studies. However, some of these initial studies were later excluded (see exclusion criteria below).

## Additional Search Strategies

Three additional strategies were employed to ensure that we identified potentially relevant reports that may not have been identified with prior searches in the reference databases: direct contact with researchers, backward searches, and forward searches. We contacted educational researchers to learn about undiscovered projects that were relevant but difficult to find, such as very recent research and unpublished reports. Specifically, we contacted researchers who had written relevant articles in the past



10 years ( $n = 4$ ). Three of the four researchers responded but did not provide additional reports. We also sent an e-mail to the American Educational Research Association Classroom Assessment Special Interest Group listserv to request that members share any research that related to SPG in the K-12 classrooms. No replies were received.

Next, we examined the reference lists of all reports that met inclusion criteria to determine whether they cited any potentially relevant reports (backward search). We then conducted a cited reference search to determine whether the reports that met inclusion criteria had been later cited by any potentially relevant reports (forward search). We reviewed these articles for relevance. For backward and forward searches, titles and abstracts were initially reviewed by the first author, and if deemed potentially relevant, the full text was obtained. Forward and backward searches were conducted on any empirical study report and/or literature review that was determined to be relevant. These two search strategies yielded 29 potential articles. Finally, we also received documents ( $n = 3$ ) from colleagues in our laboratory who were conducting searches on different but related educational research topics. The first author examined the full text of 113 studies; 53 did not pass initial screening for possible relevance according to the exclusion criteria listed above, and 27 studies required a more thorough inspection, but were ultimately excluded for the reasons listed in supplemental file B of the online supplemental materials. This resulted in a total of 33 articles eligible for the meta-analyses.

## Procedure for Synthesis of Studies

**Information retrieved from studies.** Numerous characteristics of each study were retrieved from reports and entered into a database. These characteristics encompassed six broad distinctions among studies: (a) *the research report* included basic information about the authorship and date of report appearance; (b) *study characteristics* included information about the setting, cultural context, and design features of the study; (c) *sample information* detailed the demographic characteristics of the different samples of students; (d) *grader comparison* information included general grading instructions, rubric use, and grader training; (e) *outcome measures* included details pertaining to the test format and subject; and (f) *estimate of ES* detailed the information needed to calculate an ES for the relationship between grading outcomes (e.g.,  $n$ ,  $M$ ,  $SD$ , and/or correlations). As is true in all meta-analyses, many of the study characteristics we coded either were not reported often enough or exhibited too little variability across studies to be examined through moderator analyses. Variables that were examined in the meta-analysis will be presented along with the overall results.

**Effect size estimation.** To answer the first and second research questions, we used the standardized mean difference, or Hedges'  $g$  (Hedges & Olkin, 1985), to estimate the effect of the student grader on student outcomes (i.e., subsequent learning or mean grade given) reported by each study. In most cases, we first calculated ESs (Cohen's  $d$ ; Cohen, 1988) from means and standard deviations using the ES calculator provided by Wilson (2001). We then input the ES and group  $ns$  into a statistical program (Comprehensive Meta-Analysis; Borenstein, Hedges, Higgins, & Rothstein, 2014) to calculate Hedges'  $g$ . If means and standard deviations were not available, we indirectly retrieved the information

needed to calculate  $d$ -indexes using inferential statistics (Borenstein et al., 2014; Wilson, 2001). Several reports presented separate means and standard deviations for multiple subsections of one test. In these cases, ESs were calculated for each domain.

To compare the effects of student grading on subsequent test performance, we subtracted the mean grade given by the teacher-grader group (comparison) from the mean grade given by the student-grader group (treatment) and then divided by the difference of their weighted average standard deviation. In this case, positive  $g$ -indexes indicated that the experience of student grading increased students' performance on later tests. To compare student versus teacher on grades, we subtracted the mean grade given by instructors from the mean grade given by self- or peer-graders and divided their weighted average standard deviation. Thus, positive  $g$ -indexes indicated that grades given by peers or the self were higher than grades given by instructors.

To address the third research question, for those reports contributing correlations between students and teachers, correlation coefficients were coded exactly as reported. Thus, positive correlations indicated agreement between student and expert graders and larger positive correlations indicated stronger agreement.

**Coder reliability.** Each research report was coded independently by two coders (a research assistant and a postdoctoral fellow). If there was a discrepancy in coding, the two coders discussed each disagreement until agreement was reached. If the disagreement could not be resolved, the principle investigator was consulted. Because all studies were independently coded twice and disagreements were resolved by a third independent coder, the effective reliability of codes is very high (Rosenthal, 1987) and an estimate of reliability (which would involve two new coders and an independent disagreement resolver) is not called for (APA Publications and Communications Working Group on Quantitative Research Reporting Standards, 2016).

**Identification of statistical outliers.** First, we examined the distribution of ESs, for both  $g$ -indexes and  $r$  values, to determine whether any were statistical outliers. The Grubbs (1950) test, also called "the maximum normed residual test" (also see Barnett & Lewis, 1994), identifies outliers in univariate distributions one observation at a time. If outliers were identified (using  $p < .05$ , two-tailed, as the significance level), these values were set at the value of their next nearest neighbor. Separate tests were conducted for those reports contributing  $g$ -indexes and correlations for the separate research questions. This same procedure was also applied to the distribution of samples sizes.

**Publication bias.** Despite the use of several complementary search techniques, the possibility always remains that we were unable to obtain all studies that have investigated our research questions. Therefore, we used the Duval and Tweedie (2000a, 2000b) trim-and-fill procedure to test whether the distribution of ESs used in the analyses was consistent with variation in ESs that would be predicted if the estimates were normally distributed. For example, a skewed distribution might indicate a possible publication bias created either by the study retrieval procedures or by data censoring on the part of authors. The trim-and-fill procedure provides a way to estimate the values from missing studies that need to be present to approximate a normal distribution. Often, these missing values indicate nonsignificant results that are less likely to make their way into obtainable reports.



**Independent hypothesis tests.** To avoid a potential biasing effect of multiple ESs per study, we conducted random effects meta-analyses, using robust variance estimation with small sample correction (Hedges, Tipton, & Johnson, 2010; Tanner-Smith & Tipton, 2014; Tipton, 2015). The robust variance estimator (RVE) addresses the problem of correlated group pairs by mathematically adjusting the standard errors of the ESs to account for the dependence and the small sample correction maintains appropriate Type I error rates. An intraclass correlation was specified ( $\rho = .8$ ) to estimate the ES weights. The RVE method has one important limitation: Tanner-Smith, Tipton, and Polanin (2016) assert that when is  $df < 4$  (or the number of studies  $< 5$  for a single predictor analysis), use of the RVE method is not suggested because of the unreliability of the  $t$ -distribution and underestimation of the true Type I error. In these cases, summary statistics were computed with Comprehensive Meta-Analysis software without controlling for study clustering, which is less likely to be influential with a small number of studies. Heterogeneity was assessed using  $\tau^2$ , the between-studies variance component, and the  $I^2$  statistic, which is the percentage of the total variability attributable to variation in ESs (and not sampling of participants into studies). Higher  $\tau^2$  values denote higher proportions of the observed variation to be real (Borenstein et al., 2014). Approximate guidelines for interpreting  $I^2$  values have been established at 25%, 50%, and 75% for low, medium, and large heterogeneity, respectively (Higgins & Thompson, 2002; Higgins, Thompson, Deeks, & Altman, 2003).

We first examined several study characteristics that past researchers and other scholars had suggested might be associated with SPG study outcomes as well as other important characteristics of research design. For all three meta-analyses, we examined six potential moderators that addressed differences in the classroom characteristics and SPG procedures (rubric use and training). Studies were grouped according to the (a) grade level of the students (Grades 3–8 or Grades 9–12), (b) class subject (science, technology, engineering, and math [STEM] or not-STEM), (c) rubric use, (d) whether or not students received training (no training vs. some training), (e) the length of training (less than six or more than/equal to seven exposures), and (f) mode of training (either practice or examples or multiple modes including both examples and practice).

In our analyses, each ES associated with a study was first coded as if it were an independent estimate of the relationship. Thus, we report number of reports ( $k$ ) and number of ESs (which might be more numerous) in our results.

**Software.** The Comprehensive Meta-Analysis (CMA, Version 3.3.070; Borenstein et al., 2014) software package was used to calculate the within-study variance for each study, to examine publication bias, to compute Hedges'  $g$ , and to calculate summary statistics when the number of studies was less than five. Robust variance estimation was conducted using R Package (R Core Team, 2016), with syntax provided by Tanner-Smith et al. (2016) when the number of studies to be included was more than five.

## Results

Our search strategies coupled with the inclusion/exclusion criteria identified 33 reports that represented the retrievable literature on SPG in kindergarten through 12th grade. These reports answered three unique, but related, questions on SPG. Reports were grouped into the “SPG as formative assessments” when the re-

searchers investigated the *long-term consequences* of SPG with (a) a group who participated in student grading, and (b) a group who did not (self-grading,  $k = 20$ ; peer-grading,  $k = 7$ ). Reports that compared mean grades given by teachers with those given by student-graders *on the same test* were grouped into the “SPG as summative assessments” ( $k = 9$ ). Reports that investigated the *correspondence of scores within the distribution of grades* given by student-graders and teachers were included as summative assessment but analyzed separately ( $k = 7$ ). One report tested all three questions (Sadler & Good, 2006) and another report tested the long-term effects of *both* SPG (Tseng & Tsai, 2007). These results were entered into each group of studies to which they were relevant. With multiple related outcomes appearing in some studies, a total of 86 usable  $g$ -indexes and 13 correlations were retrieved. Of note, no reports studied children younger than third grade, so our results are bound by this lower limit of generalizability. Also, no statistical outliers of sample size, Hedges'  $g$ , or Pearson's  $r$  were detected.

## SPG as Formative Assessment

This meta-analysis included reports that addressed the research question, “What are the effects of using student grading on students' subsequent performance?” A majority of the reports ( $k = 20$ ) studied the effects of self-grading on subsequent test performance (see Table 1), although a minority ( $k = 7$ ) reported on peer-grading (see Table 2). A few reports ( $k = 2$ ) contributed ESs for both self-grading and peer-grading separately; their information was included in both sets of data. Most of the studies (84%) were carried out in the United States or Canada. Sample sizes ranged from 18 to 667 students.

**Effect of self-grading on subsequent academic performance.** Table 1 summarizes the reports that examined the effect of self-grading compared with groups that had their tests graded as usual. Some reports ( $k = 14$ ) provided pretest scores for the self-grading and teacher-grading groups. We calculated Hedges'  $g$  for the pretest scores and subtracted it from the posttest  $g$ -index to compute an adjusted  $g$ -index for 30 ESs. For the six reports that did not provide a pretest score, the unadjusted  $g$ -index was used. The average report contributed more than one ES ( $M = 2.20$ ,  $SD = 1.88$ , minimum = 1, maximum = 9). Of the 44 ESs, 32 were positive (i.e., students in the self-grading condition performed better on a subsequent test than students who had not self-graded) and 12 were negative (i.e., self-grading students performed worse than comparison students). Effects sizes ranged from  $-0.82$  to  $1.75$  (see Table 1). The trim-and-fill procedure (Duval & Tweedie, 2000a, 2000b), used to test for data censoring, estimated no missing ESs from the distribution.

Using robust standard errors to account for within-study clustering based on a random effects error model, the average weighted  $g$ -index was .34, 95% confidence interval (CI) [.15, .52],  $\tau^2 = .09$ ,  $I^2 = 87.01$ .<sup>2</sup> These results suggest that, on average, self-grading in the

<sup>2</sup> We also used the independent sample as the unit of analysis and averaged across effect sizes within independent samples. The average weighted  $d$ -index was .33, 95% CI [.19, .46], for a random effects model. The test for heterogeneity of effect sizes was significant,  $Q(23) = 127.59$ ,  $p < .001$ ,  $\tau^2 = .08$ ,  $I^2 = 81.97$ , indicating that the variability in effect sizes was greater than that which would be expected because of sampling error alone.

Table 1  
*Studies Investigating the Effect of Self-Grading Skills on Subsequent Performance*

Study name	Publication type	Number of effect sizes for each study	<i>n</i>	Country	Grade	Test subject	Student training	Exposure to training	Hedges' <i>g</i>
Andrade and Boulay (2003)	J	2	119	U.S.	7 & 8	Language arts	P	2	.04 .1
Fontana and Fernandes (1994) <sup>a</sup>	J	2	667	Portugal	3 & 4	Math	E & P	6+	.15 .48
Guastello (2001) <sup>a</sup>	CP	1	167	U.S.	4	Language arts	E & P	6+	1.26
Horn (2009) <sup>a</sup>	D	1	38	U.S.	3	Language arts	E & P	1	-.21
Irwin (1973) <sup>a</sup>	D	1	266	U.S.	9–12	Mechanical drawing	ng	6+	.27
Maqsud and Pillai (1991) <sup>a</sup>	J	4	68	South Africa	9–12	Science	none	0	.11 .17 .32 .44
McDonald and Boud (2003)	J	4	515	Barbados	11	Humanities	E & P	6+	.49 .52 .27 .49
Olina and Sullivan (2004)	J	2	170	Latvia	10 & 11	Psychology	E & P	2	.14 .27
Poplin (2009) <sup>a</sup>	D	1	128	U.S.	11	History	None	6+	.25
Ramdass and Zimmerman (2008)	J	2	42	U.S.	5 & 6	Math	P	1	1.08 .35
Ross et al. (1998) <sup>a</sup>	CP	2	306	Canada	5 & 6	Math	E & P	6+	.03 -.03
Ross et al. (1999) <sup>a</sup>	J	1	296	Canada	4–6	Language arts	E & P	6+	.18
Ross et al. (2001) <sup>a</sup>	CP	9	37	Canada	11	Math	ng	1	-.23 -.82 -.58 -.54 -.37 -.06 -.27 -.15 .10
Ross et al. (2002) <sup>a</sup>	J	1	492	Canada	5 & 6	Math	E & P	6+	.38
Ross and Starling (2008)	J	3	143	Canada	9	Geography	E & P	6+	.30 .82 .38
Sadler and Good (2006) <sup>a</sup>	J	1	46	U.S.	7	Science	E & P	6+	.84
Schunk (1996) <sup>a</sup>	J	2	44	U.S.	4	Math	ng	6	.27 1.24
Wall (1982) <sup>a</sup>	J	1	44	U.S.	4	History, Spanish, reading	none	3	.00
Warner et al. (2012)	CP	1	50	U.S.	7	Math	E & P	ng	.09
Wolter (1975) <sup>a</sup>	D	3	18	U.S.	6	Language arts	P	2	.70 1.75 1.64

*Note.* D = dissertation/masters thesis; J = journal article; CP = Conference Proceedings; E = training through examples and P = training through practice.  
<sup>a</sup> Pretest and posttest scores were available and an adjusted *g*-index is reported in the table. Adjusted *g*-indices were computed by subtracting the pretest score *g*-index from the posttest *g*-index score.

classroom improved students' subsequent performance by about one third of a standard deviation compared with the performance of students who had not previously self-graded.

**Quality.** Variations in study designs are an especially important characteristic to examine when the research question involves testing a causal connection, but the research context leads to studies with less-than-ideal designs. Therefore, we first examined whether differences in study design characteristics led to differences in results. The dimensions associated with making strong

causal inferences are listed for each study in Table 2. Regrettably, the reports contained too little variation within each variable to do a formal statistical analysis. However, we could combine three of the design variables to group studies into three categories allowing different strengths of causal inference. Four reports employed random assignment, which is a principle indicator of a study's ability to draw strong causal inferences. Collectively, their average weighted *g*-index was 1.00, 95% CI [.64, 1.36],  $\tau^2 = .08$ ,  $I^2 = 42.71$ . There were also four studies that had nonequivalent control



Table 2

*Quality Indicators of Self-Grading Studies*

Study	Number of teachers (number of classes)	Type of design	A priori equated variables	Pre-/posttest design	Conditions drawn from same schools <sup>a</sup>	Statistical equating	Level of treatment assignment <sup>b</sup>	Unit of assignment same as unit of analysis
Andrade and Boulay (2003)	NR (13)	Q with student equating	Grade, achievement	No	Δ Schools (2)	Student demographics, prior achievement, grade level, school N/A	Class	No
Fontana and Fernandes (1994)	45 (45)	Q without equating	N/A	Yes	Δ Schools	N/A	Class	No
Guastello (2001)	NR (8)	T	N/A	Yes	Δ Schools (3)	N/A	Class	No
Horn (2009)	3 (3)	Q without equating	N/A	Yes	Δ Schools	N/A	Class	No
Irwin (1973)	5 (14)	Q without equating	N/A	Yes	Δ Schools (3)	N/A	Class	Yes
Maqsood and Pillai (1991)	1 (2)	Q without equating	N/A	Yes	Yes	N/A	Class	No
McDonald and Boud (2003)	20 (20)	Q without equating	N/A	No	Δ Schools (10)	N/A	Class	No
Olina and Sullivan (2004)	8 (16)	Q with school equating	Achievement	No	Δ Schools (8)	Prior achievement	School	No
Poplin (2009)	2 (4)	Q without equating	N/A	Yes	Δ Schools	N/A	Class	No
Randass and Zimmerman (2008)	NR (NR)	T	N/A	No	Δ Schools (2)	N/A	Student	Yes
Ross et al. (1998)	NR (14)	Q without equating	N/A	Yes	Yes	N/A	Class	No
Ross et al. (1999)	30 (30)	Q with teacher equating	Gender, grade, and training	Yes	Δ School districts (2)	N/A	Class	No
Ross et al. (2001)	1 (2)	Q without equating		Yes	Yes	Achievement motivation, gender, age, confidence, anxiety	Class	No
Ross et al. (2002)	24 (24)	Q with teacher equating	Grade, gender, and experience	Yes	Δ Schools	N/A	Class	No
Ross and Starling (2008)	6 (8)	Q without equating		No	Δ Schools	N/A	Class	No
Sadler and Good (2006)	1 (4)	Q without equating	N/A	Yes	Yes	N/A	Class	No
Schunk (1996)	2 (2)	T	Stratified by classroom	Yes	Yes	Ethnicity, gender	Student	Yes
Wall (1982)	NR (4)	Q without equating	N/A	Yes	Yes	N/A	Class	No
Warner et al. (2012)	NR (NR)	Q without equating		No	Yes	N/A	Class	No
Wolter (1975)	2 (2)	T	Gender	Yes	Yes	N/A	Student	Yes

*Note.* In all cases, the treatment was administered to students in the intended manner. There was no evidence of important overall or differential attrition, except in Andrade and Boulay (2003), in which the attrition was greater than 20%. NR = not reported; N/A = not applicable; Q = quasi-experiment design; T = randomized controlled trial design.

<sup>a</sup> If yes, students and classes were drawn from the same school. If no, number of different schools are given in parentheses (if reported). <sup>b</sup> In all cases, the student was used as the unit of analysis and clustering was not accounted for in the analysis. However, the tests to be graded were unique for each student (i.e., their own paper).



groups that equated the groups prior to the experimental manipulation (based on a variety of variables, e.g., grade, achievement). Their average weighted  $g$ -index was .23, 95% CI [.12, .35],  $\tau^2 = .00$ ,  $I^2 = 11.96$ . The rest of the studies had designs with nonequivalent control groups without a priori equating. Twelve studies were averaged to have a weighted  $g$ -index of .21, 95% CI [.02, .40],  $\tau^2 = .08$ ,  $I^2 = 87.63$ . Taken together, these results suggest that reports with a stronger experimental design showed larger long-term effects of self-grading, though a formal statistical test of this finding awaits future research.

There are also conclusions that can be drawn about weaknesses in the studies' designs as a collection. Most notably, all but four of the studies administered the treatment at the level of the classroom but analyzed the data using the student as the unit. Such analyses do not take intraclass dependencies into consideration. This is a failing of the studies (but one that, regrettably, occurs in many areas of classroom research).

Table C1 in supplemental file C of the online supplemental materials presents study characteristics that might influence the validity of conclusions after the data has been collected. Again, the studies did not reveal enough variation in these characteristics to allow credible statistical analyses of their influence on outcomes. These reveal that, as a group, study outcomes were likely not influenced by their level of overall or differential attrition. Also, floor and ceiling effects on pretests (when they were used) and posttests are generally not an area of concern. Three studies suggested students "volunteered." We suspect that the "no mention" studies also used such samples of convenience. Thus, although combining results across studies enhances the heterogeneity of included classrooms, the issue of how these samples of convenience (especially the use of volunteers) might differ from all classrooms remains unanswered. It also highlights the need for researchers to present more complete descriptions of their sampling procedures.

**Moderator analyses.** We conducted analyses exploring five moderators, grouped according to classroom characteristics and SPG procedures (i.e., use of rubrics and training), of the effects of self-grading on subsequent test performance. To aid in interpretation, we performed an analysis to determine whether any relationship existed between the six moderator variables. Only one such correlation is worth mentioning; perhaps not surprisingly, studies with multiple modes of training showed a significant positive relationship with length of training exposure ( $r = .76$ ,  $p = .000$ ). This correlation analysis suggests that the training variables were highly correlated with each other.

**Classroom characteristics.** Thirteen studies used students in elementary or middle school as subjects and seven studies used high school students. A moderator analysis revealed no significant difference in ESs for studies with younger compared with older students,  $t(13.5) = 1.11$ ,  $p = .29$ . Subjects were grouped according to STEM ( $k = 9$ ) subjects compared with other subjects that included language arts ( $k = 5$ ), mechanical drawing ( $k = 1$ ), humanities ( $k = 3$ ), and psychology ( $k = 1$ ). A moderator analysis showed no significant effect of self-grading outcomes for STEM classes compared with other subjects,  $t(17) = 0.08$ ,  $p = .67$ .

**Training.** Three variables captured differences in student training. First, studies were coded on whether they trained the students at all ("yes" or "no"; training presence). Only three studies provided no training to its students; most studies ( $k = 17$ )

gave some training. Thus, this moderator was analyzed. Then, studies were coded on whether they used multiple modes of training (training type; i.e., use of both practice and examples). Six studies trained through either practice or examples and 11 used both modes. A moderator analysis showed no significant effect of receiving one type of training or receiving multiple types of training,  $t(9.14) = 0.001$ ,  $p = .98$ .

Lastly, studies were grouped according to the frequency with which students self-graded (grading exposures). Studies whose students self-graded less than six times were considered to have received short-term exposure ( $k = 9$ ), and studies whose students self-graded on seven or more occasions were considered to have received long-term exposure ( $k = 10$ ). One study did not report number of grading exposures. The experience of self-grading less than six times appeared to be a natural cutoff. Specifically, studies that described a short-term intervention typically enumerated the number of occasions of self-grading compared with studies implementing self-grading practices over the course of the study/semester typically did not give an exact number; rather, these studies made self-grading an integral part of instruction. A moderator analysis did not show any effect of frequency of self-grading exposure on subsequent tests,  $t(15.11) = 0.85$ ,  $p = .41$ .

**Rubric use.** Of the 20 reports, 19 (95%) explicitly indicated that rubrics were used; Schunk (1996) made no mention of rubric use. Eleven reports described the use of specific rubrics to aid the students in the self-grading process, that is, a rubric that provided a clear set of criteria and described varying levels of quality for a specific assignment ("specific" rubric; Andrade & Valtcheva, 2009). Two studies used general rubrics in the grading process (that is, rubrics that provide some criteria reflecting the underlying skills and knowledge within the defined domain, but ultimately leaves the grader to make an overall judgment on the quality of the work) and several did not describe the rubric used ( $k = 6$ ). Only seven reports indicated the use of students to create rubrics, most of these reports (86%) originated from the same research group (Ross and colleagues).

**Effect of peer-grading on subsequent performance.** Table 3 summarizes seven reports that contributed 11 ESs for the analysis of the difference on subsequent tests between students who graded their peers and those who did not. Most of the reports used nonequivalent control groups without equating as the method of assignment ( $k = 6$ ). Of note, the one study that employed random assignment to condition demonstrated the largest effect of peer-grading ( $g = .69$ ). Generally speaking, the peer-grading studies had no serious issues with attrition or measurement ceilings and floors (see supplemental file C, table C2). Many of the studies of peer-grading were carried out in elementary and middle schools ( $k = 4$ ) and in a language arts course ( $k = 6$ ). In regard to training, almost all studies, with the exception of one (Farrell, 1977), trained their students to peer-grade and most ( $k = 5$ ) provided students with many opportunities (i.e., more than 10 peer-grading instances) to grade their peers. All of the studies reported that they had used rubrics to guide the student grading, with a majority indicating the use of specific rubrics (four of seven reports, 57%; use of general rubrics,  $k = 1$ ; information not given,  $k = 2$ ). Only one report indicated that the students aided in rubric development (Sadler & Good, 2006).

A total of seven reports with adjusted means were included in the analysis. The average report contributed more than one ES



Table 3  
*Studies Investigating the Effect of Peer-Grading Skills on Subsequent Performance*

Study name	Publication type	Number of effect sizes for each study	Type of design	Pre-/posttest design	<i>n</i>	Grade	Test subject	Student training	Exposure to training	Adjusted <i>g</i> index <sup>a</sup>
Califano (1987)	D	4	Q without equating	Yes	41	5	Language arts	E	18	.02
					48	6	Language arts	E	18	.35
										.28
Farrell (1977)	D	2	Q without equating	Yes	91	11	Language arts	None	12	−.33
									12	.32
										.27
Horn (2009)	D	1	Q without equating	Yes	37	3	Language arts	E & P	1	.45
Karegianes et al. (1980)	J	1	Q without equating	No	49	10	Language arts	P	10	.41
Pierson (1966)	D	1	Q without equating	Yes	153	9	Language arts	E	13	.15
Sadler and Good (2006)	J	1	Q without equating	Yes	73	7	Science	E & P	10	.22
Wise (1992)	D	1	T	Yes	134	8	Language arts	P	1	.69

*Note.* All students were described as being in mixed achievement classrooms, with the exception of Karegianes et al. (1980), which reported on students who were below-grade achievers. All retrieved studies were conducted in the United States. Studies were listed more than once when more than one independent sample was reported. D = dissertation/master's thesis; J = journal article; Q = quasi-experiment; T = true-experiment; E = training through examples; P = training through practice.

<sup>a</sup> Adjusted *g*-index was computed by subtracting the pretest score *g*-index from the posttest *g*-index score.

( $M = 1.57$ ,  $SD = 2.19$ , minimum = 1, maximum = 4), which ranged from  $-0.33$  to  $.69$ . Using RVE with random effects assumptions, the average weighted *g*-index of adjusted means was  $.29$ , 95% CI  $[.08, .50]$ ,  $\tau^2 = .03$ ,  $I^2 = 39.21$ . This analysis suggested that peer-grading shows a positive effect on subsequent test performance.

The trim-and-fill procedure (Duval & Tweedie, 2000a, 2000b), used to test for data censoring, estimated two missing ESs smaller than the observed mean and no evidence of missing ESs larger than the overall mean. This procedure estimated with random effects error modeling (in CMA) that the mean would decrease by  $.068$ . Thus, the analysis suggested that the observed average weighted *g*-estimate might be lower than expected had the data not been censored in some way.

In summary, studies demonstrated that both self- and peer-grading positively affected subsequent achievement performance, with self-grading showing a larger positive long-term effect. In regard to the effects of self-grading exposure, studies that implemented random assignment appeared to show larger effects than studies that did not have equivalent control groups. We found little variation in student training and rubric use; nearly all students were trained and used rubrics. We found no studies that have looked at student grading prior to third grade.

### SPG as Summative Assessment

This meta-analysis included reports that answered the research questions “How do grades assigned by students compare with grades assigned by teachers on the same outcome measure?” and “What is the degree of similarity (correlation) of a student's position on a grade distribution when grades are assigned by students and by teachers?” (see Table 4). Analyses were conducted separately for each research question. The literature search identified reports ( $k = 9$ ) with 31 ESs that compared the means of students' and teachers' assigned grades. Researchers reported ESs for self-grading ( $k = 4$ ), peer-grading ( $k = 2$ ), or both ( $k = 3$ ). The average report contributed more than one ES ( $M = 3.44$ ,  $SD =$

$2.40$ , minimum = 1, maximum = 12). Of the 31 ESs, 13 were positive and 18 were negative. A positive ES indicated that the student gave higher marks than the teacher, and a negative ES indicated that the student gave lower marks than the teacher. Sample sizes ranged from five to 184 students. A few reports ( $k = 3$ ) contributed ESs for both questions and thus were included in both sets of data. Table 4 summarizes the findings that examined mean differences between students and teachers in the grades they assigned, with self and peer *g*-indexes appearing in Columns 11 and 12, respectively.

Collectively, the studies were performed in either music classes ( $k = 4$ ) or STEM-related classes ( $k = 5$ ). Notably, all the reports originated from the United States or Taiwan. Furthermore, no studies from the United States occurred in high school, whereas most of the reports from Taiwan occurred in high school. Only one study reported that the test counted toward the final grade (Sadler & Good, 2006).

**Quality.** Weaknesses (and variation) in study designs for making causal inferences is less of an issue in research that (a) compares grades assigned by students and teachers, and (b) calculates the correlation between these assigned grades. This is because the stimulus for grading (the students' tests) and the context in which the grading takes place (the school, classroom, subject matter, etc.) are yoked for each teacher–student pair in the study. Only the grader differs between conditions. Although it might be feasible to study “peer” grading by creating experimental stimuli that vary in controlled ways (and it is not clear that such a strategy would lead to more plausibly causal inferences than allowing stimuli to vary naturally), such designs could not be used to study self-grading. Further, except for the fact that few of the studies mentioned whether students were lost from the original target population, this issues of attrition and measurement ceilings appear to be inconsequential for these studies (see supplemental file C, table C3).

One study (Sadler & Good, 2006) employed random assignment to conditions regarding whether or not the students self-graded or

Table 4  
*Studies Investigating the Difference Between Student- and Teacher-Graded Tests*

Study name	Pub type	Number of effect sizes for each study	Type of design	Country	<i>n</i>	Grade	Student type	Test subject	Student training	<i>g</i> index for self-grading	<i>g</i> index for peer-grading	<i>r</i>
Aitchison (1995)	D	1	Q without equating	U.S.	84	7, 8	M	Music	None	1.53	—	.42
Chang et al. (2012)	J	2	Q without equating	Taiwan	72	9, 10, 11, 12	M	Computer	E & P	.21	.44	.83 (self)/.28 (peer)
Davis (1981)	D	4	Q without equating	U.S.	12	5	M	Music	E & P	-.24	—	.57
										-.41	—	Ng
					5	6	M	Music	E & P	-.16	—	Ng
										-.47	—	Ng
Kruse (2006)	UW	1	Q with equating	U.S.	18	6	M	Music	E	.13	—	.63
Lin et al. (2002)	J	1	Q without equating	Taiwan	57	10	M	Engineering	None	—	-.17	.63
Sadler and Good (2006)	J	5	T	U.S.	49	7	M	Science	E	—	-.18	.91
										—	-.37	ng
					24	7	M	Science	E	—	-.18	Ng
					24	7	M	Science	E	.12	-.36	.98 (self)
Sung et al. (2005)	J	2	Q without equating	Taiwan	37	9	M	Computer Science	None	-.46	—	Ng
										-.28	—	Ng
Sung et al. (2010)	J	18	Q without equating	Taiwan	29	7	H	Music	E	-.52	-1.51	.41 (self)
					32	8	H	Music	E	-.37	-1.18	.52 (self)
					60	7	M	Music	E	-.38	-1.02	Ng
					48	8	M	Music	E	.07	-.68	Ng
					27	7	L	Music	E	.47	.69	Ng
					30	8	L	Music	E	.34	.20	Ng
Tseng and Tsai (2007)	J	3	Q without equating	Taiwan	184	10	M	Computer	None	—	.64	.71
										—	.34	.56
										—	.21	.57

*Note.* Rubrics were given to students in all studies to assist with grading. Studies were listed more than once when more than one independent sample was reported. D = dissertation/master's thesis; J = journal article; UW = unpublished work; Q = quasi-experiment; T = true experiment; M = mixed levels of student achievement; L = low achievement levels; H = high achievement levels; E = training through examples; P = training through practice; ng = not given.

peer-graded. Like all other studies, the stimuli (tests) were still yoked for students and teachers. This study reported one ES for self-grading ( $g = .12$ ) and four ESs for peer-grading (average weighted  $g$ -index =  $-.28$ ).

**Rubric use.** Of the nine reports, all indicated that rubrics were used. Most of the reports ( $k = 6$ ) described the use of general rubrics to aid the students in the grading process, whereas a small proportion of reports used specific rubrics ( $k = 3$ ). Students were not often included in rubric creation ( $k = 6$ ), with a few reports indicating student involvement in rubric creation ( $k = 3$ ).

**Student training.** Notably, four of the 9 studies did not train their students to grade. Some of the studies used examples only to train the students to grade ( $k = 3$ ), and others ( $k = 2$ ) used both examples and practice.

**Differences in mean grades assigned by self and teacher.** For the analysis examining differences between self- and teacher-grading, seven reports contributed to the summary statistic. Using RVE with random effects assumptions, the average weighted  $g$ -index of means was  $.17$ , 95% CI  $[-.41, .76]$ ,  $\tau^2 = .30$ ,  $I^2 = 89.82$ . These results suggest that, on average, primary and secondary school students assigned themselves grades that are not significantly different from the grades that teachers assigned when grading the same outcome.

The trim-and-fill procedure (Duval & Tweedie, 2000a, 2000b), used to test for data censoring, estimated two missing ESs larger than the overall mean and no evidence of missing ESs smaller than the overall mean. Specifically, the analysis (performed in CMA)

with random effects error estimated that the average would increase by  $.22$  if the missing studies were included. Thus, this analysis suggested that the estimate is lower than might have been found without publication bias or another type of data censoring (e.g., selective reporting of results by authors).

**Differences in mean grades assigned by peers and teacher.** A total of five reports contributed to the summary statistic examining the difference between grades given by peers and by teachers on the same test. Using RVE with random effects assumptions, the average weighted  $g$ -index of means was  $-.04$ , 95% CI  $[-.60, .52]$ ,  $\tau^2 = .60$ ,  $I^2 = 96.62$ . These results suggest that, on average, primary and secondary school students assigned grades to their peers that are not significantly different from the grades that teachers assigned when grading the same outcome.

The trim-and-fill procedure (Duval & Tweedie, 2000a, 2000b), used to test for data censoring, estimated with random effects modeling (in CMA) no missing ESs smaller or larger than the overall mean. Thus, this analysis suggested that the estimate approximates what might have been found without publication bias or another type of data censoring (e.g., selective reporting of results by authors).

**Distribution similarity.** There were a total of eight reports that contained 13 correlations between student-assigned and teacher-assigned grades. Sample sizes ranged from 17 to 184 students. Correlations between student graders and teacher-graders ranged between  $r = .42$  and  $.71$ . A total of six reports provided correlations between self and teacher grades, and four reports



provided correlations between peer and teacher grades (two reports contributed correlations for both self-analysis and peer-analysis).

The trim-and-fill procedure (Duval & Tweedie, 2000a, 2000b) estimated no missing ESs less than the overall mean and two ESs more than the overall mean. An estimate of the adjusted weighted overall mean difference, including the identified missing values, would increase the correlation estimate.

Using RVE with random effects assumptions, the average weighted  $r$  value for the correspondence between self-grading and teacher-grading (based on seven reports) was .67, 95% CI [.41, .93],  $\tau^2 = .05$ ,  $I^2 = 95.29$ . The trim-and-fill procedure (Duval & Tweedie, 2000a, 2000b) estimated no missing ESs less than the overall mean and two ESs more than the overall mean. The procedure, with random effects modeling in CMA, suggested that the correlation estimate would increase by .086. This procedure used for the detection of publication bias indicated that the average weighted  $r$  value is less than expected.

Based on four reports, the average weighted  $r$  value for the correlation between grades given by peers and grades given by teachers was .68, 95% CI [.32, .87],  $\tau^2 = .24$ ,  $I^2 = 97.24$ . The trim-and-fill procedure (Duval & Tweedie, 2000a, 2000b) estimated no missing ESs less than the overall mean and one ES more than the overall mean. The procedure, with random effects modeling in CMA, suggested that the correlation estimate would increase by .084, thereby indicating that the average weighted  $r$  value is less than expected. Taken together, these results suggest that, on average, primary and secondary school students assigned themselves and their peers' grades that corresponded well with the grades that instructors assigned when grading the same outcome, at least with regard to where particular students were placed in the distribution of all students.

In summary, the reports investigating SPG as summative assessment is sparse. Students in fifth to 12th grades evaluated students' tests similarly to teachers; both grades assigned and the distribution of grades showed a similarity to teacher grades. However, inferences are limited given the incomplete representation of cultures and students from primary and secondary schools.

## Discussion

This research synthesized the literature examining self-grading and peer-grading in the third through 12th grade levels using criterion-referenced testing. The survey of previous scholar writing and the meta-analyses contribute to the current body of literature by examining self-graders and peer-graders separately using criterion-referenced feedback and restricting studies to those that implemented an experimental or quasi-experimental design. This exercise helps disentangle numerous issues not separated in previous review efforts and provided formal tests for a (regrettably) few moderators that scholars and educators have posited could mediate the SPG process. It also uncovered important questions that have gone unanswered.

Our findings revealed that most of the literature surrounding SPG in primary and secondary classrooms examined SPG as formative assessment, that is, to help students do better on future tests. As expected, the practice of self-grading in the classroom showed a nontrivial effect on students' subsequent grades.

One important issue to note is the paucity of reports investigating SPG in kindergarten through second grades. Although our

meta-analyses sought to include reports containing this young group of children, no studies were found that examined any of our questions at these grade levels. Also, we did not encounter any study that investigated peer-grading in high school. Thus, we have to limit our generalizations to grading practices in the third to 12th grades for the long-term consequences of SPG, and the fifth to 12th grades for SPG as summative assessment. But these omissions may not be random. It would not be surprising if teachers feel that children 8-years-old or younger have not developed the meta-cognitive skills needed to be graders. They may also lack the emotional security to take criticism from peers (as might adolescents). For peer-grading, high school teachers may feel that the importance of grades for college admissions and the heightened role of social comparison among adolescents makes peer-grading problematic in high school. Whether these conjectures on our part actually do exist, it would be fruitful avenues for future research.

## What Are the Effects of Student Grading on Subsequent Test Performance?

Our clearest and arguably most important result is that self-grading increased academic performance on subsequent tests by about one third of a standard deviation, suggesting that active engagement in the grading process results in beneficial effects for student learning. These findings concur with and provide empirical evidence for Ross's (2006) findings from a more general review of the literature.

It is important to recognize that the self-grading condition was compared with a teacher-graded or no-grading condition (an *as usual* condition). Teachers who are prepared to provide students with instruction on grading must have a clear sense of what is important, and their instruction is likely to be reasonably well-aligned with their grading criteria. In this sense, teachers in the SPG condition are more likely to have clearer lesson objectives than those who are not.

Studies that controlled for differences between groups through random assignment showed larger effects on the long-term effect of self-grading. In addition, this finding is consistent with other similar meta-analyses in SPG research at the college level (Falchikov & Boud, 1989; Falchikov & Goldfinch, 2000), which suggested effects of methodological quality (based on various experimental-level variables in addition to completeness in reporting of information) on the reported outcomes of SPG research.

Given that training of students to self-grade is often espoused as an essential ingredient of successful self-grading, it is surprising that our indicators of training (presence, length, and type) did not moderate the effects of self-grading on student achievement. However, all of the studies provided students with rubrics and most provided training, which suggests a high degree of student supervision and structure present throughout the studies. This scaffolding of instruction to students may have diminished our ability to find effects of training because of low variability between studies. It does point out that training is viewed as a critical element to the success of SPG and suggests that future studies that experimentally manipulated the presence, amount, and type of training would be highly informative.

Similar to the self-grading experience, the peer-grading experience also benefited subsequent test performance (even when we used a robust procedure for estimating effects). However, the small



number of studies contained in the peer-grading literature limits the generalizability of the findings to the U.S. population. Although the current results contributed additional studies on peer-grading and limited studies to ones that used criterion-referenced assessments, its findings are similar to Topping's (2013) review that suggested that although the situation is slowly changing, more studies are needed that investigate peer-grading in primary and secondary classrooms.

Importantly, the effect of peer-grading was smaller compared with the effect of self-grading on subsequent test performance. This finding suggests that self-grading may affect metacognitive skills (e.g., reflection and internalization) more than peer-grading among third through 12th graders. Self-grading relies on a student's metacognitive competencies by drawing on self-observation, self-judgment, task analysis, self-control, and so forth (Brown & Harris, 2013). Self-grading may improve student academic performance by teaching students to use and rehearse metacognitive skills. Self-regulation is related to academic achievement: students that are capable of setting goals, making flexible plans to meet them, and monitoring their progress are more likely to perform better in school than students who are not (Andrade & Valtcheva, 2009), as evidenced by improvement in self-grading skills over time (Butler & Lee, 2010). Furthermore, self-regulation skills (setting goals, deliberating about strategies, managing motivation) are argued to be the most useful skill for students to be effective learners (Butler & Winne, 1995).

With regard to the theoretical rationales for SPG writ large, the existing studies did not directly measure mediating variables suggested by the relevant theories. For example, based on theoretical explanations of SPG's effect on learning, the students exposed to SPG should show increases in their sense of autonomy, self-monitoring, and sense of fairness in grading. These have never been measured in SPG studies, but the results were consistent with the theoretical predictions. It would be both interesting and informative if future studies of SPG included measures of these variables.

### **Do Mean Grades Differ When Students or Teachers Are the Graders?**

The meta-analysis investigating student and teacher mean grade comparisons found little difference between grades assigned by students and teachers. Self-grading means were in the predicted direction ( $g = .17$ ), but not significantly so, perhaps related to the lack of power to test this effect. Peer-assigned grades hardly differed at all from teacher grades ( $g = -.04$ ). This meta-analysis perhaps serves to somewhat allay teachers' conception that students are unable to grade themselves and others without grade inflation; instead, it appears that students in primary and secondary classrooms give scores similar in mean to scores given by teachers.

Meta-analyses investigating the same research questions at the college level showed that students graded tests between one fourth to one half of a standard deviation higher than teachers (Falchikov & Boud, 1989; Falchikov & Goldfinch, 2000). Thus, it appears that the difference between students and teachers emerges as grades become more consequential. First, these findings may stem from more supervision and structure in the grading process for younger students, as all reports in the primary and secondary school meta-analysis provided rubrics for the student to grade.

Second, more competition exists as students move through school and grades have longer term and more direct implications for continuing education, which might lead to a greater pressure to grade oneself with more leniency and their peers with more rivalry (Sebba et al., 2008). Third, students in earlier grades may be more inclined to explicitly follow teachers' instructions. College students are typically given more autonomy in the academic atmosphere, whereas primary and secondary school students typically experience more direct guidance from their teachers. Lastly, tests at higher grade levels contain more complexity, which might require a greater level of inference and/or metacognitive abilities (Hovardas et al., 2014). Future research might directly investigate the differences in SPG implementation as students move through levels of schooling. Taken together, our synthesis confirms that developmental and contextual differences exist in the implementation of SPG in the college versus the primary and secondary school settings.

Importantly, most of the reports did not have the test count toward the final grade, thereby making the student-grading process a low-stake activity in the classroom. Also, no studies from the United States occurred in high school, whereas most of the reports from Taiwan occurred in high school.

### **What Is the Correlation Between Student Graders and Teachers?**

The grades assigned by middle and high school students demonstrated a moderate relationship with those assigned by teachers in regard to the placement of students within grade distributions ( $r = .67$ ). This finding suggests that about 45% of the variance in grades was shared by students and teachers. Notably, the correlations were all in the positive direction, which suggests generally good correspondence between students and teachers. It also suggests that future research should examine what explains the remaining variance in both student and teacher-graded scores. Although a stronger emphasis on effort reflected in grades given by students is often proposed (Stipek, 1981; Stipek & Tannatt, 1984), we know of no study that directly investigates this supposition or whether its influence diminishes over across development. Taken together, moderate correlations and good mean correspondence between student and teacher grades indicates that SPG can be successful as a summative assessment in primary and secondary classrooms, though the weight given to test grades in determining a student's overall grades is yet to be tested.

### **SPG as Formative or Summative Assessment**

More studies investigated the effect of SPG as formative assessment, thereby giving students the opportunity of revision and improvement over the long term. Studies of SPG in primary and secondary classrooms rarely implemented SPG as summative assessment, which involves students grading their own or others' work for a final grade. This imbalance of literature is perhaps reflective of the general movement to emphasize formative assessment and reflects the degree of difficulty in using students as the judge for a final grade, especially when students are very young or preparing to apply for college (when grades count the most). This finding may also be because of the pervasive practice of teacher-controlled summative results in the kindergarten through 12th



grade classrooms. Understandingly, it appears to be difficult for precollege teachers to report grades judged by students and/or their peers to stakeholders and parents because of concerns about how reliable and similar these may be to teacher judgments. However, the current meta-analysis suggested that fifth through 12th graders showed relatively similar grading outcomes compared with teachers.

## Limitations

Our syntheses are confined by the typical limitations implicit in the nature of meta-analyses. Specifically, our analysis was limited by the level of completeness and specificity of reporting found in the studies that we identified through our literature search. Issues of statistical power restricted our investigation of many moderators, such that numbers were small or unequally distributed between moderator variable groupings.

Also, our methods would have been strengthened if we were able to use intraclass coefficients to control for clustering effects within classes. However, the studies reported within these meta-analyses did not report intraclass correlations, and estimated values were not available that would have approximated our studies based on similar sampling strategies, populations, and outcome measures (Hedges & Rhoads, 2011).

## Future Research

A more complete picture of SPG is needed in primary and secondary schools. In addition to the research directions mentioned above, future research might investigate the development of SPG skills in kindergarten through the third grade. We understand, however, that SPG at these grades may raise concerns about the students' emotional reactions and its impact on self-concept, especially among poorer performers. Additionally, research has suggested that younger children's ability to engage in SPG may be compromised because of their strong emphasis on effort when grading (Nicholls, Patashnick, & Mettetal, 1986; Stipek & Tannatt, 1984). Thus, we recommend SPG in young children with great caution.

Notably, many of the indicators suggested many of the studies included in the meta-analyses had design weaknesses, for example, experimental and control groups were often drawn from different (nonequivalent) schools and the unit of assignment rarely was the unit of statistical analysis. Although the occasional use of random assignment and the consistency of findings is encouraging, these weaknesses suggest that school-based research looking at SPG needs more rigorous tests of effectiveness.

In addition, these meta-analyses were not able to address how student-level variables might affect the degree to which SPG is beneficial. For example, more studies are needed that investigate how a student's level of ability influences the effectiveness of SPG. Furthermore, studies that address the influence of students' grade level and class subject matter on SPG outcomes are critical to establish a more comprehensive understanding of SPG and achievement.

## Conclusion

Our meta-analyses do provide fresh data and insights indicating the following:

- Primary and secondary students demonstrate enhanced learning in the future when they have previously self- or peer-graded. These results suggest that when students partake in SPG, they may develop clearer retention and/or understanding of the assessed material.
- Students can self-grade and peer-grade relatively similarly to teachers. When using SPG as formative evaluation, the process of SPG is important for the learning experience and the actual importance of the grade is diminished.

Thus, SPG could effectively be implemented more frequently in the third through 12th grade classrooms. Self-grading as formative assessment appears to be the most favorable way to practice SPG in the classroom. Additionally, it appears that teachers (and researchers) understand that they must give students training and support in order for students to benefit from the procedures. Training of students represents a significant initial time commitment on the part of the teacher; however, any new curriculum endeavor typically requires a similar initial effort.

## References

References marked with a single asterisk (\*) indicate studies included in the formative assessment meta-analysis; references marked with a double asterisk (\*\*) indicate studies included in the summative assessment meta-analyses.

- Aitchison, R. E. (1995). *The effects of self-evaluation techniques on the musical performance, self-evaluation accuracy, motivation, and self-esteem of middle school instrumental music students*. Iowa City, Iowa: University of Iowa.
- \*Andrade, H. G., & Boulay, B. A. (2003). Role of rubric-referenced self-assessment in learning to write. *The Journal of Educational Research*, 97, 21–30. <http://dx.doi.org/10.1080/00220670309596625>
- Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education*, 32, 159–181. <http://dx.doi.org/10.1080/02602930600801928>
- Andrade, H., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice*, 27, 3–13. <http://dx.doi.org/10.1111/j.1745-3992.2008.00118.x>
- Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice*, 48, 12–19. <http://dx.doi.org/10.1080/00405840802577544>
- APA Publications and Communications Working Group on Quantitative Research Reporting Standards. (2016). *Journal article reporting standards for quantitative research in psychology—once again*. Manuscript under review.
- Atkinson, K. M., Sanchez, C. E., Koenka, A. C., Moshontz, H., & Cooper, H. (2016). *Who makes the grade? A synthesis of research comparing self, peer and instructor grades in college classrooms*. Manuscript submitted for publication.
- Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, 61, 213–238. <http://dx.doi.org/10.3102/00346543061002213>
- Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (Vol. 3). New York, NY: Wiley.
- Beach, R. (1979). The effects of between-draft teacher evaluation versus student self-evaluation on high school student's revising of rough drafts. *Research in the Teaching of English*, 13, 111–119.
- Berger, R., Rugen, L., & Woodfin, L. (2014). *Leaders of their own learning: Transforming schools through student-engaged assessment*. Hoboken, NJ: Wiley.



- Bickmore, D. K. (1981). *The effects of student self-evaluation and pupil-teacher conferences on student perceptions, self concepts, and learning*. Provo, UT: Department of Secondary Education and Foundations, Brigham Young University.
- Bonniol, J. J. (1981). Influence de l'explicitation des critères utilisés sur le fonctionnement des mécanismes d'évaluation d'une production scolaire [Influence of the exploitation of the criteria used on the mechanisms of a scholastic evaluation]. *Bulletin de Psychologie*, 353, 173–186.
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2014). Comprehensive meta-analysis (version 3.0 and Version 3.3.070). Englewood, NJ: Biostat.
- Boud, D. (1991). *Implementing student self-assessment*. Hammondville, New South Wales: Higher Education Research and Development Society of Australasia.
- Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education*, 10, 161–180. [http://dx.doi.org/10.1207/s15324818ame1002\\_4](http://dx.doi.org/10.1207/s15324818ame1002_4)
- Brown, G. T., & Harris, L. R. (2013). *Student self-assessment*. Thousand Oaks, CA: Sage. <http://dx.doi.org/10.4135/9781452218649.n21>
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65, 245–281. <http://dx.doi.org/10.3102/00346543065003245>
- Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, 27, 5–31. <http://dx.doi.org/10.1177/0265532209346370>
- Califano, L. Z. (1987). *Teacher and peer editing: Their effects on students' writing as measured by t-unit length, holistic scoring, and the attitudes of fifth and sixth grade students*. Flagstaff, AZ: Northern Arizona University.
- \*\*Chang, C.-C., Tseng, K.-H., & Lou, S.-J. (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a web-based portfolio assessment environment for high school students. *Computers & Education*, 58, 303–320. <http://dx.doi.org/10.1016/j.compedu.2011.08.005>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge.
- Crocker, A., & Cheeseman, R. (1988). The ability of young children to rank themselves for academic ability. *Educational Studies*, 14, 105–110.
- Davis, L. M. (1981). *The effects of structured singing activities and self-evaluation practice on elementary band students' instrumental music performance, melodic tonal imagery, self-evaluation, and attitude*. Columbus, OH: The Ohio State University.
- Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education*, 24, 331–350. <http://dx.doi.org/10.1080/03075079912331379935>
- Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5, 69–106. <http://dx.doi.org/10.1111/j.1529-1006.2004.00018.x>
- Duval, S., & Tweedie, R. (2000a). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95, 89–98.
- Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455–463. <http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x>
- Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, 41, 1040–1048. <http://dx.doi.org/10.1037/0003-066X.41.10.1040>
- Evans, E. D., & Engelberg, R. A. (1988). Student perceptions of school grading. *Journal of Research & Development in Education*, 21, 45–54.
- Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, 59, 395–430. <http://dx.doi.org/10.3102/00346543059004395>
- Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70, 287–322. <http://dx.doi.org/10.3102/00346543070003287>
- Farrell, K. J. (1977). *A comparison of three instructional approaches for teaching written composition to high school juniors: Teacher lecture, peer evaluation, and group tutoring*. Boston, MA: Boston University.
- Fitzgerald, J., & Markham, L. R. (1987). Teaching children about revision in writing. *Cognition and Instruction*, 4, 3–24. [http://dx.doi.org/10.1207/s1532690xcic0401\\_1](http://dx.doi.org/10.1207/s1532690xcic0401_1)
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34, 906–911. <http://dx.doi.org/10.1037/0003-066X.34.10.906>
- \*Fontana, D., & Fernandes, M. (1994). Improvements in mathematics performance as a consequence of self-assessment in Portuguese primary school pupils. *British Journal of Educational Psychology*, 64, 407–417. <http://dx.doi.org/10.1111/j.2044-8279.1994.tb01112.x>
- Goodrich, H. (1996). *Student self-assessment: At the intersection of meta-cognition and authentic assessment*. Cambridge, MA: Harvard Graduate School of Education.
- Graham, S., Harris, K., & Hebert, M. (2011). *Informing writing: The benefits of formative assessment. A report from Carnegie Corporation of New York*. New York, NY: Carnegie Corporation of New York.
- Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing. *The Elementary School Journal*, 115, 523–547. <http://dx.doi.org/10.1086/681947>
- Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics*, 21, 27–58. <http://dx.doi.org/10.1214/aoms/1177729885>
- \*Guastello, E. F. (2001). Parents as partners: Improving children's writing. In W. M. Linek, E. G. Sturtevant, J. A. R. Dugan, & P. E. Linder (Eds.), *Celebrating the voices of literacy: Yearbook of the College Reading Association* (pp. 279–295). Readyville, TN: College Reading Association.
- Hedges, L., & Olkin, I. (1985). *Statistical models for meta-analysis*. New York, NY: Academic Press.
- Hedges, L. V., & Rhoads, C. H. (2011). Correcting an analysis of variance for clustering. *British Journal of Mathematical and Statistical Psychology*, 64, 20–37. <http://dx.doi.org/10.1111/j.2044-8317.2010.02005.x>
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1, 39–65. <http://dx.doi.org/10.1002/jrsm.5>
- Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21, 1539–1558. <http://dx.doi.org/10.1002/sim.1186>
- Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, 327, 557–560. <http://dx.doi.org/10.1136/bmj.327.7414.557>
- Hodgson, C. (2010). Assessment for learning in science: What works? *Primary Science*, 115, 14–16.
- \*Horn, G. C. (2009). *Rubrics and revision: What are the effects of 3rd graders using rubrics to self-assess or peer-assess drafts of writing?* Boise, ID: Boise State University.
- Hovardas, T., Tsivitanidou, O. E., & Zacharia, Z. C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education*, 71, 133–152.
- Ikeguchi, C. B. (1996). *Self Assessment and ESL Competence of Japanese Returnees*. Tsukuba Women's University, Tsukuba, Japan. Retrieved from <http://files.eric.ed.gov/fulltext/ED399798.pdf>
- Irwin, J. L. (1973). *An investigation of the effects of student self-evaluation and marking on achievement of cognitive and affective objectives of a*



- basic mechanical drawing course*. State College, PA: Pennsylvania State University.
- \*Karegianes, M. L., Pascarella, E. T., & Pflaum, S. W. (1980). The effects of peer editing on the writing proficiency of low-achieving tenth grade students. *The Journal of Educational Research*, 73, 203–207. <http://dx.doi.org/10.1080/00220671.1980.10885236>
- Klenowski, V. (1995). Student self-evaluation processes in student-centred teaching and learning contexts of Australia and England. *Assessment in Education: Principles, Policy & Practice*, 2, 145–163. <http://dx.doi.org/10.1080/0969594950020203>
- Koivula, N., Hassmén, P., & Hunt, D. P. (2001). Performance on the Swedish Scholastic Aptitude Test: Effects of self-assessment and gender. *Sex Roles*, 44, 629–645. <http://dx.doi.org/10.1023/A:1012203412708>
- \*\*Kruse, N. B. (2006). *The effect of instruction on sixth grade band students' abilities to self-rate etude performance*. East Lansing, MI: Michigan State University.
- Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, 75, 63–82. <http://dx.doi.org/10.3102/00346543075001063>
- Lane, S. (2012). Performance assessment. In J. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 313–330). Thousand Oaks, CA: Sage.
- \*\*Lin, S. S., Liu, E. Z., & Yuan, S. (2002). Student attitudes toward networked peer assessment: Case studies of undergraduate students and senior high school students. *International Journal of Instructional Media*, 29, 241–254.
- \*Maqsd, M., & Pillai, C. M. (1991). Effect of self-scoring on subsequent performances in academic achievement tests. *Educational Research*, 33, 151–154. <http://dx.doi.org/10.1080/0013188910330208>
- \*McDonald, B., & Boud, D. (2003). The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examinations. *Assessment in Education: Principles, Policy & Practice*, 10, 209–220. <http://dx.doi.org/10.1080/0969594032000121289>
- Nicholls, J. G., Patashnick, M., & Mettetal, G. (1986). Conceptions of ability and intelligence. *Child Development*, 57, 636–645. <http://dx.doi.org/10.2307/1130342>
- \*Olina, Z., & Sullivan, H. J. (2004). Student self-evaluation, teacher evaluation, and learner performance. *Educational Technology Research and Development*, 52, 5–22. <http://dx.doi.org/10.1007/BF02504672>
- Paquette, K. R. (2008). Integrating the 6 + 1 writing traits model with cross-age tutoring: An investigation of elementary students' writing development. *Literacy Research and Instruction*, 48, 28–38. <http://dx.doi.org/10.1080/19388070802226261>
- \*Pierson, H. (1967). *Peer and teacher correction: A comparison of the effects of two methods of teaching composition in grade nine English classes*. New York, NY: New York University.
- \*Poplin, B. D. (2009). *Effects of student self-corrective measures on learning and standardized test scores*. Lynchburg, VA: Liberty University.
- \*Ramdass, D., & Zimmerman, B. J. (2008). Effects of self-correction strategy training on middle school students' self-efficacy, self-evaluation, and mathematics division learning. *Journal of Advanced Academics*, 20, 18–41.
- Rosenthal, R. (1987). *Judgment studies: Design, analysis, and meta-analysis*. New York, NY: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511527807>
- Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research & Evaluation*, 11, 1–13.
- \*Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2001). *Self-evaluation in grade 11 mathematics: Effects on achievement and student beliefs about ability*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
- \*Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5–6 mathematics effects on problem-solving achievement. *Educational Assessment*, 8, 43–58. [http://dx.doi.org/10.1207/S15326977EA0801\\_03](http://dx.doi.org/10.1207/S15326977EA0801_03)
- \*Ross, J. A., Rolheiser, C., & Hoaboam-Gray, A. (1998). *Impact of self-evaluation training on mathematics achievement in cooperative learning environment*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
- \*Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing*, 6, 107–132. [http://dx.doi.org/10.1016/S1075-2935\(99\)00003-3](http://dx.doi.org/10.1016/S1075-2935(99)00003-3)
- \*Ross, J. A., & Starling, M. (2008). Self-assessment in a technology-supported environment: The case of grade 9 geography. *Assessment in Education: Principles, Policy & Practice*, 15, 183–199. <http://dx.doi.org/10.1080/09695940802164218>
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78. <http://dx.doi.org/10.1037/0003-066X.55.1.68>
- \*/\*\*Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational Assessment*, 11, 1–31. [http://dx.doi.org/10.1207/s15326977ea1101\\_1](http://dx.doi.org/10.1207/s15326977ea1101_1)
- \*Schunk, D. H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal*, 33, 359–382. <http://dx.doi.org/10.3102/00028312033002359>
- Sebba, J., Crick, R., Yu, G., Lawson, H., Harlen, W., & Durant, K. (2008). *Systematic review of research evidence of the impact on students in secondary schools of self and peer assessment. I. Research Evidence in Education Library series*. London, UK: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London. Retrieved from <http://eppi.ioe.ac.uk/cms/Default.aspx>
- Stipek, D. J. (1981). Children's perceptions of their own and their classmates' ability. *Journal of Educational Psychology*, 73, 404–410. <http://dx.doi.org/10.1037/0022-0663.73.3.404>
- Stipek, D. J., & Tannatt, L. M. (1984). Children's judgments of their own and their peers' academic competence. *Journal of Educational Psychology*, 76, 75–84. <http://dx.doi.org/10.1037/0022-0663.76.1.75>
- \*\*Sung, Y.-T., Chang, K.-E., Chang, T.-H., & Yu, W.-C. (2010). How many heads are better than one? The reliability and validity of teenagers' self- and peer assessments. *Journal of Adolescence*, 33, 135–145. <http://dx.doi.org/10.1016/j.adolescence.2009.04.004>
- \*\*Sung, Y.-T., Chang, K.-E., Chiou, S.-K., & Hou, H.-T. (2005). The design and application of a web-based self-and peer-assessment system. *Computers & Education*, 45, 187–202. <http://dx.doi.org/10.1016/j.compedu.2004.07.002>
- Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, 5, 13–30. <http://dx.doi.org/10.1002/jrsm.1091>
- Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology*, 2, 85–112.
- R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Tillema, H., Leenknicht, M., & Segers, M. (2011). Assessing assessment quality: Criteria for quality assurance in design of (peer) assessment for learning—a review of research studies. *Studies in Educational Evaluation*, 37, 25–34. <http://dx.doi.org/10.1016/j.stueduc.2011.03.004>

- Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, 20, 375–393. <http://dx.doi.org/10.1037/met0000011>
- Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, 68, 249–276. <http://dx.doi.org/10.3102/00346543068003249>
- Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. J. R. C. Dochy, & E. Cascalla (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 55–87). New York, NY: Springer.
- Topping, K. (2009). Peer assessment. *Theory Into Practice*, 48, 20–27. <http://dx.doi.org/10.1080/00405840802577569>
- Topping, K. (Ed.). (2013). *Peers as a source of formative and summative assessment*. Thousand Oaks, CA: Sage. <http://dx.doi.org/10.4135/9781452218649.n22>
- \*\*Tseng, S.-C., & Tsai, C.-C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education*, 49, 1161–1174. <http://dx.doi.org/10.1016/j.compedu.2006.01.007>
- van Gennip, N. A., Segers, M. S., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review*, 4, 41–54. <http://dx.doi.org/10.1016/j.edurev.2008.11.002>
- van Kraayenoord, C. E., & Paris, S. G. (1997). Australian students' self-appraisal of their work samples and academic progress. *The Elementary School Journal*, 97, 523–537. <http://dx.doi.org/10.1086/461879>
- van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, 20, 270–279. <http://dx.doi.org/10.1016/j.learninstruc.2009.08.004>
- \*Wall, S. M. (1982). Effects of systematic self-monitoring and self-reinforcement in children's management of test performances. *The Journal of Psychology: Interdisciplinary and Applied*, 111, 129–136. <http://dx.doi.org/10.1080/00223980.1982.9923524>
- Wang, M. C., Haertel, G. D., & Walberg, H. J. (1990). What influences learning? A content analysis of review literature. *The Journal of Educational Research*, 84, 30–43. <http://dx.doi.org/10.1080/00220671.1990.10885988>
- \*Warner, Z. B., Chen, F., & Andrade, H. (2012). Student Self-Assessment in Middle School Mathematics: A Pilot Study. Proceedings of the Northeastern Educational Research Association Conference 2012, 5. Retrieved from [http://digitalcommons.uconn.edu/nera\\_2012/5](http://digitalcommons.uconn.edu/nera_2012/5)
- Wilson, D. (2001). *Effect size determination program*. College Park, MD: University of Maryland.
- Wise, W. G. (1992). *The effects of revision instruction on eighth graders' persuasive writing* (Unpublished doctoral dissertation). College Park, MD: University of Maryland.
- Wright, C. R., & Houck, J. W. (1995). Gender Differences among self-assessments, teacher ratings, grades, and aptitude test scores for a sample of students attending rural secondary schools. *Educational and Psychological Measurement*, 55, 743–752.
- Wolter, D. R. (1975). *Effect of feedback on performance on a creative writing task* (Unpublished doctoral dissertation). Ann Arbor, MI: University of Michigan.

Received August 30, 2015

Revision received December 18, 2016

Accepted December 26, 2016 ■

UNITED STATES POSTAL SERVICE (All Periodicals Publications Except Registered Publications)

Journal of Educational Psychology

Issue No. 48

Frequency: Jan, Feb, Apr, May, Jul, Aug, Oct, Nov

Subscription Rates: \$120 (US), \$150 (Foreign)

Postmaster: Send address changes to: American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242

Second-Class Bulk Rate: 750 First Street, NE, Washington, DC 20002-4242

Postage paid at Washington, DC and at additional mailing offices.

PSN 0022-0671

UNITED STATES POSTAL SERVICE (All Periodicals Publications Except Registered Publications)

Journal of Educational Psychology

Issue No. 48

Frequency: Jan, Feb, Apr, May, Jul, Aug, Oct, Nov

Subscription Rates: \$120 (US), \$150 (Foreign)

Postmaster: Send address changes to: American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242

Second-Class Bulk Rate: 750 First Street, NE, Washington, DC 20002-4242

Postage paid at Washington, DC and at additional mailing offices.

PSN 0022-0671

UNITED STATES POSTAL SERVICE (All Periodicals Publications Except Registered Publications)

Journal of Educational Psychology

Issue No. 48

Frequency: Jan, Feb, Apr, May, Jul, Aug, Oct, Nov

Subscription Rates: \$120 (US), \$150 (Foreign)

Postmaster: Send address changes to: American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242

Second-Class Bulk Rate: 750 First Street, NE, Washington, DC 20002-4242

Postage paid at Washington, DC and at additional mailing offices.

PSN 0022-0671



# Four Semesters Investigating Frequency of Testing, the Testing Effect, and Transfer of Training

Donald J. Foss and Joseph W. Pirozzolo  
University of Houston

We carried out 4 semester-long studies of student performance in a college research methods course (total  $N = 588$ ). Two sections of it were taught each semester with systematic and controlled differences between them. Key manipulations were repeated (with some variation) across the 4 terms, allowing assessment of replicability of effects. Variables studied included frequency of tests (e.g., 2 vs. 8 in-class exams), the repetition of some and not other exam items (i.e., the testing effect), and variation of test items between the in-class exams and the final exam (e.g., identical items vs. controlled changes in items). Some studies also manipulated presence or absence of low-stakes quizzes. The repetition of test items generally led to better performance. However, we did not observe consistent superiority for items that were repeated exactly over those that were repeated in modified form; the reverse was more often the case. The effect of the low-stakes quizzes was minimal at best. Results are discussed in terms of memory and transfer of training models.

## *Educational Impact and Implications Statement*

Can we find inexpensive and easily adaptable modifications to teaching methods that positively impact student outcomes? These studies provide a positive answer to that question. The work is based on laboratory findings that frequent tests and frequent attempts to recall the same material (1) aid learning and memory, and (2) help students apply what they've learned to new problems. The present studies took place in large-enrollment college classes across four semesters. Within each semester two sections of an undergraduate course were taught in a highly similar fashion, primarily differing in the number of tests given and whether items that appeared on an earlier test were repeated on the final exam. In addition, some of the repeated items were identically so, while other 'repeated' items tested the same concepts but with different wording. We found evidence that frequent testing and repetition of tested items can improve course performance up to about 10%, though the results varied across the studies so further work is needed to clarify why. We also observed that under some circumstances students did as well or even better on re-worded test items as they did when the item was repeated in exactly the same words.

**Keywords:** testing effect, frequency of testing, transfer of training, college learning

Among the oft-studied variables that affect learning and retention are: the frequency of tests; whether, how often, and in what form the material has previously been tested, that is, the testing effect; and manipulations that affect the extent to which learning transfers to new test environments. The present work assesses these and related effects across four entire semesters in a college course. That is, it presents a study and three variations (near replications) of it using highly ecologically valid materials and settings in both high-stakes and (in some instances) low- or no-stakes testing. One major motivation for these studies is to further determine whether the laboratory-based findings associated with these variables generalize to the college classroom over an entire

course. Another is to examine some boundary conditions on their effectiveness and even on their proposed mechanisms.

## The Testing Effect

The testing effect has a substantial history. More than a century ago Myers (1914) reported a study involving seventh and eighth graders who were given a surprise recall test after getting a list of 10 words to spell. Some students who got a delayed-recall test, for example an hour after spelling the words, had also been given an immediate recall test without feedback. On the delayed test these students recalled more items than those who had not been given

This article was published Online First March 23, 2017.

Donald J. Foss and Joseph W. Pirozzolo, Department of Psychology, University of Houston.

Particular thanks to David Francis for extensive consultation on the statistical analyses reported here. Thanks, also, to the following lab instructors whose assistance was invaluable in this work: Melissa Trevino, Aurora Ramos-Núñez, Nikki Arrington, Emily Barton, Maya Ravid-Greene, Catherine Jockell, and Katy Reynolds. We also appre-

ciate help from Yusra Ahmed and Amelia Coffman, and from Randolph Bias who made many useful suggestions on how to make this paper better. Finally, we are grateful for the assistance of the University of Houston Registrar, Debbie Henry, and her collegial staff, especially Tracie Briscoe.

Correspondence concerning this article should be addressed to Donald J. Foss, Department of Psychology, University of Houston, 126 Heyne Building, Houston, TX 77204-5022. E-mail: dfoss@uh.edu

the immediate recall test. Taking the first test helped performance on the later one: hence, “the testing effect.”

Myers also tells of a student (it was a different time: a footnote informs us it was Miss Margaret Griffith) who had recited two prose passages before the College Literary Society, one 464 words long and the other 1,242 words. He asked her to recite the longer one to him once each week for 7 weeks, without feedback and with no further study or practice. At the end of that time she recited it perfectly. He then asked her to recall the shorter passage and reported that she could recall less than half of it—a case study of the testing effect. Myers (1914) put his conclusion succinctly: “Simple recall of stimuli wholly or partly learned aids in their retention” (p. 128).

Myers was clearly onto something. While its prominence has sometimes waned, as evidenced by the title of Glover’s (1989) paper: “The ‘Testing’ Phenomenon: Not Gone But Nearly Forgotten,” the evidence for it is substantial and has continued to grow (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Kang, McDaniel, & Pashler, 2011; Rawson & Dunlosky, 2013; Roediger, & Karpicke, 2006a; Rowland, 2014). Work on the testing effect has been further extended to the classroom, including studies using grade school and high school students, and testing materials drawn from lessons in science (e.g., McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013), history (e.g., Carpenter, Pashler, & Cepeda, 2009; McDermott, Agarwal, D’Antonio, Roediger, & McDaniel, 2014), social studies (e.g., Roediger, Agarwal, McDaniel, & McDermott, 2011), and others. There also has been some research with college (and even medical) students testing the testing effect (e.g., Cranney, Ahn, McKinnon, Morris, & Watts, 2009; Kromann, Jensen, & Ringsted, 2009; McDaniel, Roediger, & McDermott, 2007). See also Roediger and Karpicke (2006b).

In their comprehensive summary, Dunlosky et al. (2013) say, “. . . we rate practice testing as having high utility” (p. 35). And in the Pashler et al. (2007) report the recommendation to apply the testing effect within the nation’s classrooms (recommendation 5b) is one of only two said to have a strong level of evidence in its favor (*italics in original*):

5. Use quizzing to promote learning. *Use quizzing with active retrieval of information at all phases of the learning process to exploit the ability of retrieval directly to facilitate long-lasting memory traces. . . . 5b. Use quizzes to reexpose students to key content.* (p. 2)

### The Frequency of Testing

The frequency of testing also has a substantial history. More than 80 years ago, Keys (1934) carried out an early study investigating the frequency of testing using the complex materials from an actual college course. He examined the effects of weekly versus monthly tests on retention in a course on educational psychology. To do so, he taught two sections of the same course, in the same lecture hall at the University of California, “. . . and great pains were taken to keep the instruction identical” across them (Keys, 1934, p. 429). To quickly crush any budding sense of nostalgia for the cozy classrooms of the past, we learn that enrollment in the two sections totaled three hundred sixty students. Keys used the same test items in each section, one group getting more frequent and shorter exams than the other. The in-course exams comprised

true-false and completion items in the ratio of 7 to 1, while the final exam was composed of 100% true-false items.

Aside from posting grades, no feedback was provided to students on their in-course exam performance, although persistent students could see their corrected exams. They had to be persistent because, “the time and place were intentionally made so inconvenient” (Keys, 1934, p. 430) that, on average, only about 10% of the students succeeding in seeing an exam. (Again, it was a different time.)

Students in the weekly exam section did significantly better (by 12%) on the in-course exams themselves than did students in the monthly exam section. However, this may have been due to the facts that the weekly exams (a) covered less material, and (b) were administered closer to the time the course material was presented. More importantly, Keys actually gave two “final” exams in parallel forms: one surprise exam on the penultimate class day, and two weeks later the announced final. On the unannounced final the weekly exam group significantly outperformed the monthly exam group—they differed by 7%. That difference is unlikely due to a difference between the two classes in the time between presentation and test given that Keys attempted to keep the instruction identical between them. In contrast, there was no difference between the weekly and the monthly groups on the announced final, which Keys suggests might have been due to “cramming.”

Since the early work of Keys (1934) there have been hundreds of studies on frequency effects, most conducted in the laboratory. Others have been carried out with relatively simple educational materials (e.g., foreign language vocabulary), and following Keys, a few others have employed more complex course content. With some exceptions and reservations (e.g., Donovan & Radosevich, 1999; Ross & Henry, 1939), most investigations have found that more tests lead to better retention (e.g., Bahrack, Bahrack, Bahrack, & Bahrack, 1993; Bangert-Drowns, Kulik, & Kulik, 1991; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Gaynor & Millham, 1976; Kika, McLaughlin, & Dixon, 1992; Leeming, 2002). Thus, while there is no doubt the effect is real, some investigators (e.g., Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Delaney, Verkoeijen, & Spiguel, 2010) note that when carefully analyzed the apparently simple pattern of results can actually be quite complicated.

The “more is better” generalization has been refined, for example by Bangert-Drowns et al. (1991) whose meta-analysis of frequent classroom testing showed that with each successive test the additional effect size shrinks. They also made the useful observation that effect size differences between frequent and less frequent numbers of tests depend upon the absolute number of tests in the “less frequent” group, with the difference between zero and one in-course exam being quite substantial. Indeed, the difference between “no midterm” and “one midterm” was so great that the present authors considered it ethically dubious to conduct a classroom-based experiment at the limit; that is, one in which the less frequent group has no midterm. See also Cepeda et al. (2008).

To date there has been a relatively small (though growing) subset of studies that followed Keys (1934) by carrying out the investigation over an entire course or substantial parts of one (see below). A number of questions remain about the conditions under which one will see frequency effects, what variables affect effect sizes, what interactions might be practically important, whether and what types of test materials matter and why, and so forth.



In a heroic review of the (probable) practical efficacy of 10 learning techniques derived from cognitive and educational psychology (Dunlosky et al., 2013), the authors give high marks to the likely positive influence of distributed practice (and by inference to frequent testing). They conclude that, “distributed practice should work for complex materials as well.” However, in what we do not believe is a proforma statement, they quickly add, “Future research should examine this issue” (Dunlosky et al., 2013, p. 40). Similarly, in a report sponsored by the National Center for Education Research (Pashler et al., 2007), the first recommendation, one said to have a moderate level of evidence in support of it, is (*italics in original*): “Space learning over time. *Arrange to review key elements of course content after a delay of several weeks to several months after initial presentation*” (p. 2). Notable for our present purpose, the authors also go on to say, “One limitation of the literature is that few studies have examined acquisition of complex bodies of structured information” (Pashler et al., 2007, p. 6). The present work is intended to help limit the extent of that limitation.

### Testing and Transfer

In the present studies we addressed another question of both theoretical and practical importance, namely: does an increase in retrieval probability due to earlier testing require presenting the identical question on subsequent test(s), and if not, what determines whether the student will recognize that the question is interrogating the same conceptual knowledge and will thereby benefit from the earlier test? The literature has mixed results on this topic suggesting that there may not be a simple function relating test items when their form varies across two or more administrations.

As an example of a problematic issue, in a laboratory experiment using a chapter from a biology textbook, Wooldridge, Bugg, McDaniel, and Liu (2014) only found an advantage due to prior testing when the final test items were both based on factual material and identical to items tested in the first presentation. They found no advantage for items that required applying the facts, even when the items were identical in the two tests. Similarly, they did not see a testing effect when the item on the second test was “related” to the first test item, whether it was a factual or an application item.

And recently, Nguyen and McDaniel (2015) report on a laboratory experiment using published course materials such as test-bank questions. They note that “when quiz and test items are haphazardly sampled” the type and degree of relationships between those items vary. In that case they “found no net gain on a final exam for students who took the quizzing program compared with those students who were instructed to highlight while studying” (Nguyen and McDaniel, 2015, p. 89; highlighting being a common control condition in laboratory studies on the testing effect). The overall null effect was apparently due to observing the expected advantage of quizzing when the items tested the same concept in the same format, and observing a reverse testing effect when the quiz items asked about a new example of a previously tested concept. They dubbed this among the “ugly” findings, and cautioned (p. 89) that if items “are haphazardly sampled, teachers must be cautious in assuming that testing will confer benefits for exam performance” (Nguyen and McDaniel, 2015, p. 89).

Even after all this time, then, neither the testing effect, the frequency of testing effect, or the transfer effect stories are over. After stipulating their enthusiasm for the testing and frequency effects, Dunlosky and Rawson (2012) go on to say: “Despite the promise of these techniques, however, further research is needed to more firmly establish their efficacy in the classroom and to discover how they can best be used to ensure robust learning and comprehension” (p. 254).

### Testing in the College Classroom

Our current work examines aspects of the testing effect, the frequency of testing, and transfer effects in situ. We carried out a series of studies in the tradition of Keys (1934) and a few others (e.g., Bangert-Drowns et al., 1991; Carpenter et al., 2009; Gaynor & Millham, 1976; Leeming, 2002; Pennebaker, Gosling, & Ferrell, 2013); that is, research carried out in classrooms over entire courses (or close to it, e.g., Mawhinney, Bostow, Laws, Blumenfeld, & Hopkins, 1971; Roediger et al., 2011a), and therefore based upon complex and interrelated materials. Such an approach has the disadvantage of complexity in terms of teasing out the causes of observed effects. But it also has the advantage of being an extension and “conceptual” replication of laboratory studies—as well as of other class-based work; and of having a relatively short generalization path to applications if the results warrant.

Though we have not mentioned it previously, one practical motivation for this work was to determine whether relatively small changes in course structure can lead to measureable and meaningful changes in student learning and performance. If so, then perhaps that can help convince colleagues to embrace those changes. To put it another way, we are interested in finding evidence-based, inexpensive, scalable, and easily adoptable and adaptable modifications to teaching methods that positively impact student outcomes.

### Common Framework and Methods for the Research

Over each of four consecutive semesters (not including summer courses) the senior author taught two sections of a college course called Methods in Psychology, a required course for both psychology majors and minors at the University of Houston (UH). Though recommended for sophomores, each section had enrollees from that level to those graduating at the end of the term. The course covered an introduction to scientific methods in psychology, numerous aspects of experimental design and interpretation, a basic introduction to descriptive statistics and null-hypothesis decision-making, and ethics in research. It also introduced related topics such as base rates and psychological effects on decision-making.

Within the constraints of the various test schedules described below, and following Keys (1934), “great pains were taken to keep the instruction identical” between the lecture sections each term—more honestly, to keep it highly similar given the differences in students’ questions and something akin to the “personality” of each class. In each semester the students in the two lecture sections saw the same PowerPoint material (typically also made available on a web site on the day of the lecture  $\pm 1$ ), were given the same homework problems via the web site, were presented with the same “ripped from the headlines” topic at the start of almost every class in order to demonstrate and stimulate discussion of the



ubiquity and content of claims about human behavior and how they might be tested, heard the same (what the instructor liked to consider) jokes, and got the exact same exam items both on the in-class midterm exams and on the final exam. We followed the university's final exam schedule each semester, which typically meant it was administered about a week to 10 days after the last class meeting.

Early in all sections of the course throughout this project the instructor emphasized the importance of active learning, in general, and self-testing, in particular, and posted that advice as part of an exam hint document.

For practical reasons we were not able to randomly assign students to the two sections. In order to help make the student characteristics similar across them, one section was taught at 10 a.m. and the other at 11:30 a.m. on Tuesdays and Thursdays. Each class lasted 60 min. Across the two years we varied whether the 10 a.m. or the 11:30 a.m. class was in the "frequent exam" condition. As will be further described below, we also obtained standardized test and GPA information about each student so that we could regress out certain aptitude/accomplishment scores to yield more comparable results between the courses each semester.

In addition to a lecture section, each student was also in a smaller "laboratory" or "recitation" section that met for an hour each week. These meetings focused on learning to use a university library, American Psychological Association (APA) writing style, and developing a research proposal. Graduate student instructors taught these sections following a common syllabus. Performance in the "lab sections" was evaluated separately and was largely determined by a paper proposing an experimental study. Evaluations in those sections used a common rubric, and we made an effort to have similar grading standards across sections in any given semester. While the lab grade was factored into the student's final grade in the course, it had no impact on the scores made by students on the in-class midterms and final exams that we report here. With an important exception, noted below, material from the lecture section was not covered in the laboratory meetings, and therefore we will report only on data drawn from the lecture sections.

### Common Methods, Procedures, and Design of Materials

Although there were substantial conceptual and operational similarities across the semesters, we did make changes as we proceeded—including, in the second year, adding a potentially important variable that may have impacted the results, and making a significant procedural change in Study 4. Accordingly, we will describe the changes made in each study in the Method sections below.

In each of the four studies we compared student performance across two sections, one of which—the "standard testing" class—had two midterms, while the "frequent testing" class was given four midterms in Study 1 and eight midterms in the following three studies. The students were not aware that we referred to them as the frequent and standard classes; each student had access to the syllabus for his or her respective class and was aware of how many tests would occur throughout the semester for that class. By the end of the semester each class was given the same midterm test items, and the two sections each took the same final exam. After each midterm (in one of the next two class meetings) the instructor

went over the exam in class showing the keyed correct answer and responded to questions. In addition, students were invited to visit the teaching assistant to look over their exam(s) and the key. They were not, however, allowed to copy the exam. In these large classes only a small minority of students availed themselves of this opportunity.

The simple hypothesis is that the frequent testing class will outperform the standard testing class on the common final exam, (as well as on the total points earned across all exams). There are a number of reasons for this simple prediction. For example, frequent testing likely leads to more studying as well as more spaced study of a subset of the materials. And that, in turn, likely leads frequently tested students to make more attempts at retrieving relevant information. The present studies were not primarily designed to examine this last issue—though they do look at a related one.

We also asked whether material tested on one of the midterm exams would lead to an increase in retrieval probability for that material on the final exam; and, if such an advantage exists, whether it is necessary that the later test present the item in identical format. In each study the final exam repeated some items from a midterm exam, thereby allowing a direct test of the testing effect over a semester-long course. In addition, in Studies 1, 2, and 3 we also systematically varied the form of some final exam questions relative to earlier ones, as follows: on the midterm exams we used a mixture of multiple-choice (MC) and short-answer (SA) items (equal numbers of each). Two forms of each midterm were developed: the items that appeared in MC format on one form were in SA format on the other. Approximately half of each class was consistently given each form. The final exam also had two forms built on the same principle: SA items on one form occurred as MC items on the other. Importantly, on the final exam we exactly repeated a subset of items from the midterms and "flipped" others such that if it appeared in MC format on a midterm it was rewritten to be an SA item on the final, and vice versa. We will dub it the MC-MC condition when an MC item from one of the midterms also appeared in MC format on the final exam, and the MC-SA condition when an MC item from one of the midterms appeared on the final exam in SA format. SA-SA, and SA-MC refer to analogous conditions when the item appeared in SA format on one of the midterms. The flipped items allow a simple test of whether the retested item must be identical in form to receive an advantage in a course context. The final exam also included items not previously tested.

To be more specific about the form of the final exams in Studies 1–3, each contained 64 items of which 32 (16 MC, 16 SA) were new to the students. The remaining 32 were constructed as follows: Of the 16 MC items, eight were exact duplicates of items from an earlier test: four from the exam(s) given in the first half of the class, and four from the exam(s) given in the second half. Thus, these were the MC-MC items. The other eight MC items were flipped versions of SA items from the earlier exams (thus, SA-MC): again, four from the exam(s) given in the first half of the class and four from the second. An analogous procedure was used with the remaining 16 SA items, eight being exact duplicates from earlier tests (SA-SA), and eight being flipped versions of items previously seen in MC format (MC-SA).

In summary, each of the first three studies examined the testing effect (testing itself aids learning) and the frequency of testing in



a college course. As well, each of these studies examined whether repeated items had to be in the same format (MC or SA) on each administration to experience the advantage of prior testing. We expect the section with more frequent exams to do better than the standard two-exam section, the repeated items (including the flipped ones) to do better on the final than nonrepeated ones, and to observe a greater advantage for the exact items than for the flipped ones. The fourth study tested the same variables, but manipulated the construct associated with testing the transfer effect in a different way (see below).

**Common aspects of the participants.** Each term the participants were undergraduate students at the sophomore level and above who were enrolled in one of two sections of a Methods in Psychology course at the UH. Nearly all of them took the course because it is required for psychology majors and minors. At this time UH has among the most diverse college student populations in the country (e.g., there is no ethnic majority on campus), and psychology courses reflect that diversity. A preponderance of UH psychology majors are females (as is typical across the United States at this time), and the same is true for participants in these studies.

In addition to collecting their exam scores, we obtained for each student certain standard data collected by the university. In the ideal case those would be the same data for each participant, but at UH some students present with SAT scores, fewer with ACT scores, and some have neither because in this state a student can transfer to a senior level institution if he or she meets certain criteria, for example, has successfully completed a set of core courses at a community college or other public 4-year school. The UH gets a lot of transfer students and does not collect SAT or ACT scores from most of them. We also obtained cumulative GPA scores for each student. These GPAs were based on varying numbers of courses—those an individual student had completed at UH up to and including the semester in which he or she took the class.

Because we could not randomly assign students to class sections, we constructed an “aptitude” score for each participant based upon the standardized test and GPA information we obtained. For each of the following on which an individual had a score, SAT Math and Critical Thinking, ACT, and UH GPA, we computed the student’s standard score relative to all students in all eight sections across the two years. If a student had a score on both SAT measures, we averaged those standard scores to get one for SAT. Then we averaged the standard scores to obtain a single value for each participant’s aptitude.

Over the course of this work we made three possibly significant changes in methods, and a few that we consider minor ones. These will be described below.

**Common analyses.** In each of the following four studies, we conducted two main analyses. First, to investigate the effects of prior testing on final exam performance, a repeated measures analysis of covariance (ANCOVA; SAS 9.4) was performed using the mean score on each of the six types of final exam items for each participant as the dependent variable (for clarification, these six types of items were MC-MC, SA-MC, SA-SA, MC-SA, new MC, and new SA in Studies 1–3; a similar design using six types of items was used in Study IV). Although the factors included in the models vary across the four studies, each analysis included the following factors: student aptitude, testing frequency, and item type. This analysis provides tests of the effect of repetition (the testing effect), transfer of learning (performance on flipped and

new items in Studies 1–3, and related and new items in Study 4), and potential interactions between testing frequency and item type. We report results from fixed effects and orthogonal contrasts in this analysis to evaluate our specific hypotheses in each study. In Studies 1–3 orthogonal contrasts are used to test (a) whether students perform better on repeated (MC-MC, SA-SA, SA-MC, and MC-SA) than “new” (new MC and new SA) items, and (b) whether students perform better on “exact” (MC-MC and SA-SA) than flipped (SA-MC and MC-SA) items.

Second, to perform a more direct test of the effect of testing frequency, an ANCOVA (SAS 9.4) on the final exam score was used. Factors included in these models also varied across the studies; however, in each study the model included student aptitude and testing frequency.

## Study 1

To recap, Study 1 was designed to test whether students in a college course perform better on final exam items taken during a previous exam and whether taking more frequent tests benefits subsequent final exam performance. Students in two sections of a college Methods course either took two or four course exams during a 15-week semester, and then took the same comprehensive final exam. Items on the final exam either (a) appeared exactly as taken during one of the course exams, or (b) asked the same question as a previous item but in a different form (MC or SA), or (c) were not previously tested in any form. We predicted that students would perform better on items that they had previously seen (in any form) than on new items, and better on exact items than flipped ones. We also predicted that students in the more frequently tested class would perform better on the final exam overall.

## Method

**Participants.** There were 75 students in the standard class and 84 in the frequently tested one.

**Materials and procedure.** Two sections of the Methods in Psychology three-credit course were taught. The standard testing section (two midterms and two pop quizzes) was offered at 10 a.m. on Tuesdays and Thursdays, while the frequent testing section (four midterms, four pop quizzes) took place at 11:30 a.m. on those same days. There were 29 class meetings across 15 weeks.

The two midterms in the standard class took place in Weeks 6 and 12. Each exam contained 20 items: half MC and half SA. The four midterms in the frequent class took place in Weeks 3, 5, 9, and 12. Each of those exams had 10 items, half MC and half SA. In total, each class was given the same 40 midterm items. Two forms of the test were used in each class as described earlier.

The pop quizzes contained three to five MC items and were unannounced. They were included to stimulate and reward attendance. Questions were presented via PowerPoint slides, with students marking and then turning in answer sheets. After that, the instructor presented the correct answer for each item. Pop quizzes in the frequent class occurred in Weeks 2, 5, 7, and 14; while in the standard class they were given in Weeks 5 and 14. The final exam was constructed as described above in the Common Methods section.

In Study 1 we encountered missing data on item-by-item performance on the final exam for 39 participants. However, we retained the total final exam score for each of these participants.



For this reason, the item type analysis uses 39 fewer participants than the final exam score analysis.

## Results

Results from a repeated measures ANCOVA revealed significant main effects of student aptitude,  $F(1, 117) = 16.84, p < .0001$ , and item type,  $F(5, 117) = 45.51, p < .0001$ . The main effect of testing frequency,  $F(1, 117) = 0.63, p = .43$ , and the interaction of testing frequency and item type,  $F(5, 117) = 1.72, p = .14$ , were not significant predictors of final exam performance. A specific contrast revealed that students performed significantly better on repeated, that is, both exact and flipped items (least squares  $M = .72$ ; a model adjusted mean, henceforth "LS  $M$ ") than on new items (LS  $M = .63$ ),  $F(1, 117) = 57.59, p < .0001$ . This exact versus new item contrast yield a Cohen's  $d$  effect size of .60.<sup>1</sup> No significant difference was found between exact (LS  $M = .71$ ) and flipped items (LS  $M = .70$ ;  $d = .15$ ). Figure 1 shows performance on the final exam for the standard and frequent classes by item type.

An additional analysis to test the effect of testing frequency on final exam scores (excluding item characteristics) again found a significant effect of student aptitude,  $F(1, 155) = 68.12, p < .0001$ . The effect of testing frequency was not significant,  $F(1, 155) = .11, p = .74$ ; comparable performance was observed between the frequent (LS  $M = .66$ ) and standard (LS  $M = .65$ ) classes ( $d = .07$ ). The interaction between testing frequency and student aptitude,  $F(1, 155) = .37, p = .55$  was not significant.

## Discussion

Students performed significantly better on items that had appeared on a previous exam. Interestingly, this repetition effect did not depend on whether the repeated items were in the same form as previously seen, as evidenced by the finding that students did not perform significantly better on exact items than on flipped items. This finding differs from that of Nguyen and McDaniel (2015), though it replicates other results in the literature (Bjork, Little, & Storm, 2014) and suggests that testing during college exams may facilitate a near form of transfer of learning.

The effect of testing frequency was not significant in Study 1, contrary to our hypotheses. Over an entire semester, the increase

from two to four exams may not be enough to make a difference in exam performance. So one reason for Study 2 was to increase the number of tests in the frequent condition. We also modestly increased the number of midterm exam items in order to sample topics more broadly and to give some additional power to the comparison of item types (e.g., identical vs. flipped).

## Study 2

Study 2 was designed to see whether increasing the number of in-course exams would replicate the direction of findings from Study 1 and to further examine the effect sizes. In this semester (and in Studies 3 and 4) the frequent testing group received eight short in-course exams while the standard testing group again got two. Increasing the number of announced exams almost certainly increases the number of times the typical student reviews the material, and likely increases the number of times that students make relevant retrieval efforts while preparing for and taking those exams. And, of course, the shorter, more frequent exams generally mean that the amount of material to be reviewed is less. However, in this study the students were informed that later exams could ask about material from earlier in the course—and they were reminded that the final exam was comprehensive. In fact, a small number of items on the later in-course exams did ask about material covered earlier.

## Method

**Participants.** Class sizes were considerably smaller in Study 2: there were 34 students in the standard exam class and 36 in the frequently tested one.

**Materials and procedure.** In Study 2 the frequent testing section took place at 10:00 a.m. on Tuesdays and Thursdays, while the standard testing section was offered at 11:30 a.m. those same days. There were 28 class meetings across 14 weeks, with a 1-week break after the eighth week.

The two midterms in the standard class occurred in Weeks 6 and 12. Each of these exams contained 24 items, half MC and half SA. The eight in-course exams in the frequent class took place in Weeks 2, 3, 4, 5, 7, 8, 9, and 11. Each of those exams had six items, half MC and half SA. Overall, both classes were given the same 48 midterm items. The final exam was created as described above. No pop quizzes were given in this semester.

We report results for students who finished the course and who took both midterms if in the standard class, and who took at least seven of the eight exams if in the frequent class. For the three students who missed one of the eight, we assigned their average score on the remaining seven to the missing score for the purpose of calculating a fair cumulative score. In Study 2 we used analogous analytic strategies to those performed in Study 1.

## Results

The LS means on proportion correct are shown in Figure 2. Study 2 provided similar results to those in Study 1. An analysis of final exam performance on the various types of items showed significant

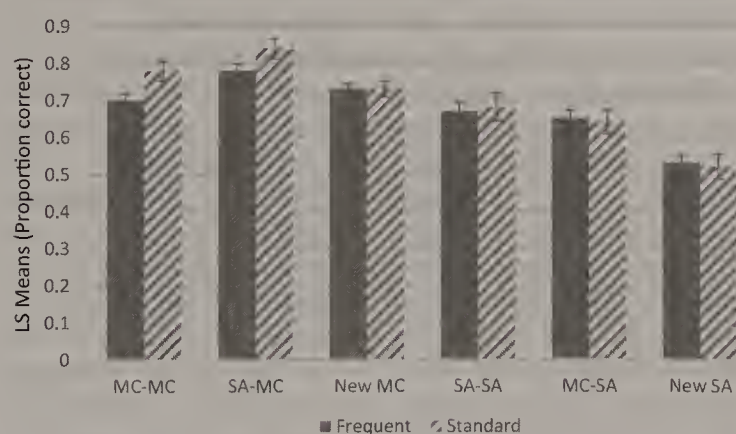


Figure 1. Least squares means (proportion correct) on the final exam in Study 1 as a function of testing frequency and item type. Error bars represent standard errors of the means.

<sup>1</sup> Cohen's  $d$  for the final exam outcome was calculated by dividing the difference in LS means between the two groups of interest by the pooled within-group standard deviation.



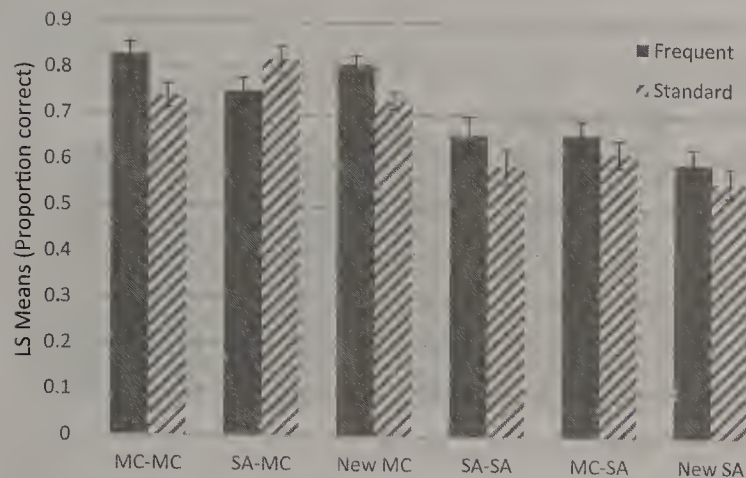


Figure 2. Least squares means (proportion correct) on the final exam in Study 2 as a function of testing frequency and item type. Error bars represent standard errors of the means.

effects of student aptitude,  $F(1, 67) = 30.43, p < .0001$ , and item type,  $F(5, 67) = 32.26, p < .0001$ . The effect of testing frequency,  $F(1, 67) = 1.87, p = .15$ , was not statistically significant, though it appeared to be stronger in Study 2 than Study 1 despite fewer participants. Interestingly, we observed a significant interaction between testing frequency and item type,  $F(5, 67) = 3.92, p = .004$  (see Figure 2). As was also shown in Study 1, students performed better on repeated items (LS  $M = .71$ ) than on new ones (LS  $M = .68; d = .20$ ),  $F(1, 67) = 7.83, p = .0007$ ; while the performance difference on exact (LS  $M = .71$ ) versus flipped items (LS  $M = .71; d = -.05$ )<sup>2</sup> was not significant,  $F(1, 67) = .25, p = .62$ .

An ANCOVA on the final exam score showed a significant effect of student aptitude,  $F(1, 66) = 44.82, p < .0001$ , but no significant effect of testing frequency (frequent LS  $M = .71$ ; standard LS  $M = .68; d = .20$ ),  $F(1, 66) = 2.12, p = .15$ , or the interaction between testing frequency and student aptitude,  $F(1, 66) = .56, p = .46$ .

## Discussion

In Study 2 we again found evidence for the testing effect: students performed better on repeated items than on new ones. And, as also found in Study 1, the repetition effect did not differ between items that were identical to those from a previous exam and those that were flipped. While there was not an overall effect for test frequency, we note that frequently tested students were correct on a higher proportion of items in five of the six item types, including the new items that they had not previously seen. The only major change in design from Study 1 to Study 2 was the increase in number of exams for the frequent class from four exams to eight. As noted, although the effect of testing frequency was not significant in Study 2, it trended stronger than in Study 1 even though the sample size (and thus the power of the comparison) was substantially smaller.

## Study 3

## Method

In Study 3 the basic testing frequency manipulation was again two versus eight in-class exams. However, in this semester we added a new component to the manipulation in order to examine

and compare the effects of low-stakes as well as high-stakes testing. Roediger et al. (2011b) report that sixth grade students improved their performance on graded free-recall exams after receiving low-stakes MC quizzes two days earlier. On an end-of-semester MC test the students also performed better on previously quizzed items than on nonquizzed ones. In their study low stakes was equivalent to no stakes because no grade depended upon the quiz performance.

Study 3 builds on this work by presenting a subset of students with no-stakes quizzes over a semester. We should quickly note that Study 3 does not use their manipulation nor directly address the testing effect because none of the no-stakes quiz items appeared later on the final exam. In this study the quiz dates were not announced in advance so students in the quiz sections could not specifically prepare for them. However, we presume that they did make additional efforts to recall aspects of the course material during the quizzes themselves. Is that enough to produce better performance on the final? One reason it might be is that, unlike studies involving lists, a course involves interrelated material such that both reading about and making retrieval attempts for Concept A may also help the later retrieval of a (related) Concept B.

**Participants.** Seventy students were in the standard class, of whom 44 received the low-stakes quizzes and 26 did not. Seventy-five students were in the frequent class; of those, 47 were administered the quizzes and 28 were not.

**Materials and procedure.** In the current study we took advantage of the eight separate laboratory sections to present a subset of students from each lecture section with a series of six short quizzes spaced across the semester. Two of the five lab sections in the standard class, and two of the three in the frequent class, were arbitrarily chosen to administer the quizzes, which were given in a “pop quiz” fashion—that is, unannounced. Each of the six quizzes contained six items, three MC and three SA. They were presented in PowerPoint format, and the instructor gave feedback on the correct answers immediately after collecting the answer sheets. While the quiz items asked about lecture material, they did not duplicate any actual exam items. Students in the quiz sections were informed that the quizzes, while collected, were not going to be graded—that they were just for practice and for their benefit. They were, however, told that they would receive extra credit for being present and taking the quizzes.

The quizzes were administered in lab sections prior to Exams 1, 2, 3, 5, 7, and 8 for the frequent exam group; and three of the six quizzes were given prior to Exam 1 for the standard group.

To summarize this change, the addition of the quiz variable meant that we had four between-subjects groups in Study 3 (and similarly in Study 4) determined by crossing the frequency of midterm variable with the quiz versus no-quiz variable. The midterms were high-stakes exams—performance on them affected the students’ grades—while the quizzes were low- (or no-) stakes because quiz performance had no effect on grades. This  $2 \times 2$  design led to cells with 2 (high stakes), 8 (high stakes), 8 (2 high stakes, 6 no stakes), and 14 (8 high stakes, 6 no stakes) entries. The quiz items themselves were drawn from the course material, but none were given on the exams in the lecture section. Thus, the

<sup>2</sup> We report a negative effect size when the observed difference was opposite to the hypothesized effect.



quizzes added some possibly relevant additional tests, but did not contribute to our study of the testing (repetition) effect. Student quiz participation statistics (for students in quiz sections) in Studies 3 and 4 are shown in Table 1. Though students took a higher percentage of the quizzes in Study 3 as compared with Study 4, in both studies a large percentage of students took at least three or more quizzes (approximately 90% in Study 3 and 60% in Study 4).

Study 3 also modified the composition of the exam items in order to allow cleaner and more interpretable comparisons of the testing effect across item types. Consider the comparison on the final exam between MC-MC items and SA-MC items—that is, when we compare performance of MC items on the final exam when preceded either by (a) the identical MC item on a midterm, or (b) by the same item presented in SA form on the midterm, that is, the flipped items. Call them MC<sub>I</sub> and MC<sub>F</sub>, respectively. In order to tighten this comparison, we examined the performance on final exam items in the previous semesters and constructed the final exam in Study 3 such that prior performance on MC<sub>I</sub> and MC<sub>F</sub> were equated (i.e., within ±1% in each pair). Any performance difference due to inherent item difficulty is thereby controlled. We carried out an analogous process for SA-SA and MC-SA items such that the SA<sub>I</sub> and SA<sub>F</sub> items were equated within ±4% in each pair.

In this study the standard testing section was offered at 10:00 a.m. on Tuesdays and Thursdays, while the frequent testing section took place at 11:30 a.m. on those same days. There were 29 class meetings across 15 weeks.

The two midterms in the standard class took place in Weeks 6 and 12. Each contained 24 items, half MC and half SA. The eight in-course exams in the frequent class took place in Weeks 2, 4, 5, 7, 8, 9, 11, and 13. Each had six items: half MC and half SA. Overall, both the standard and the frequent classes were given the same 48 midterm items. Final exams were constructed as described earlier. Analyses performed in Study 3 were comparable to analyses performed in previous studies, while including “quiz status” as a term in our model to test the effect of taking no-stakes quizzes and its interaction with other factors.

Results

The pattern of results on the final exam for Study 3 is shown in Figure 3. The analysis of item performance revealed significant effects of student aptitude,  $F(1, 140) = 65.32, p < .0001$ , item type,  $F(5, 140) = 41.37, p < .0001$ , and a significant interaction between testing frequency and item type,  $F(5, 140) = 3.17, p < .01$ . Testing frequency,  $F(1, 140) = 1.72, p = .19$ , quiz status,  $F(1,$

Table 1  
Mean Quizzes Taken and Percentage of Students Taking Each Quiz in Studies 3 and 4

Number of quizzes taken	Study 3	Study 4
Mean quizzes taken	4.33	3.26
One quiz (%)	4	18
Two quizzes (%)	7	22
Three quizzes (%)	17	15
Four quizzes (%)	24	20
Five quizzes (%)	20	9
Six quizzes (%)	28	16

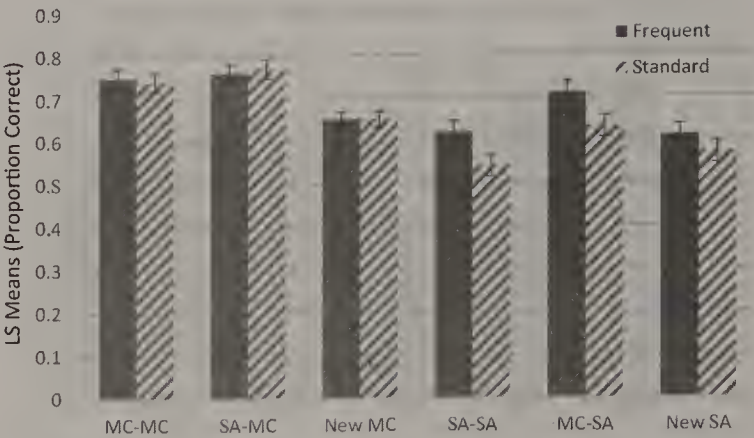


Figure 3. Least squares means (proportion correct) on the final exam in Study 3 as a function of testing frequency and item type. Error bars represent standard errors of the means.

140) = .13,  $p = .72$ , and the interactions between testing frequency and quiz status,  $F(1, 140) = 1.19, p = .28$ , and between quiz status and item type,  $F(5, 140) = .12, p = .99$ , were not significant. We again found that students performed better on repeated (LS  $M = .69$ ) than new items (LS  $M = .62; d = .39$ ),  $F(1, 140) = 53.16, p < .0001$ . Contrary to our hypothesis we found that students performed better on flipped items (LS  $M = .72$ ) than exact items (LS  $M = .66; d = -.33$ ),  $F(1, 140) = 22.14, p < .0001$ .

The LS means for each combination of quiz status (0, 6) and exam frequency (S = standard = 2; F = frequent = 8) from Study 3 are shown in Table 2. When the results of the final exam are analyzed, the overall effect due to the frequency of in-class exams (frequent LS  $M = .68$ ; standard LS  $M = .63; d = .28$ ) was statistically significant  $F(1, 140) = 4.12, p = .04$ . In contrast, on the final exam the main effect of quiz status (quiz LS  $M = .65$ ; LS  $M = .66; d = -.06$ ),  $F(1, 140) = .13, p = .72$ , and the interaction of testing frequency and quiz,  $F(1, 140) = .54, p = .46$ , was not significant.

Discussion

Study 3 again found evidence consistent with the testing effect in a college course: previous exposure to exam items improved performance on the final exam. It also found significantly superior performance of the frequent versus the standard class on the final exam. An inspection of these data hints that the difference between the frequent and standard exam schedules may be greater for SA items than for MC items, suggesting that frequent testing improves recall memory performance more than recognition memory. When inspecting Figure 3 it is evident that, although frequent and standard classes do not differ on final exam MC items (MC-MC, SA-MC, and new MC), the frequent class consistently outperforms the standard class on SA items (SA-SA, MC-SA, and new SA). That led us to make a methodological change in Study 4.

In contrast, there was no hint that taking the low-stakes quizzes led to an increase in exam performance. We mentioned earlier some aspects of our design that might have militated against observing such an effect. These include its no-stakes feature and the fact that they were unannounced. Perhaps most importantly, the quiz items were not identical with final exam items, nor did



Table 2  
*LS Means (SD) on Cumulative Score by Testing Frequency and Quiz Condition for Study 3*

Quiz condition	Frequent	Standard	Mean
Quiz	.69 ( $\pm$ .17)	.62 ( $\pm$ .18)	.65
No quiz	.68 ( $\pm$ .18)	.65 ( $\pm$ .13)	.66
Mean	.68	.63	

*Note.* Table shows the least squares (LS) means for cumulative score in each cell of the Testing Frequency  $\times$  Quiz interaction in Study 3.

they as closely tap the same conceptual information as did, for example, the flipped items on the exams. However, it seems highly likely that students did make recall efforts on the quizzes, and that when presented feedback about the correct responses when finished with a quiz they would have made an attempt to update information about the course material. Even so, we found no evidence that students generalized anything they may have learned from the quizzes to the items on the final exam.

#### Study 4

Study 3 found more consistent evidence for the testing effect among the MC items than the SA items. That is, both the MC-MC and SA-MC conditions were superior to the new MC condition, while the SA-SA condition was nearly identical to the new SA condition. Also, we observed a somewhat larger frequency effect overall for SA items than for MC ones. In Study 4 we omitted MC items to examine whether the repetition and (especially) the frequency effects would be amplified if all test items were presented in the more difficult recall format, one that may require more robust retrieval efforts. (In this study we also added 8 items at the end of the final exam to explore another hypothesis. We will not report those data here.)

Study 4 also included no-stakes SA quizzes given in some of the lab sections. We conjectured that there would be some effect from the no-stakes quizzing done in the labs, again due to the more process-intense activities needed to respond to the SA quiz items. Finally, in Study 4 the students in the quiz sections were told in advance when the quizzes would occur. Thus it was possible for them to prepare for the quizzes (even though they were told that the quizzes did not count toward their course grade).

With respect to the testing effect, Study 4 also systematically manipulated the relationship between midterm items and final exam items in a new way. Rather than changing items from MC to SA format and vice versa, in Study 4 we varied the degree of relationship between certain items on a midterm and on the final. Three types of relationships were used: (a) some final exam items were exact duplicates of midterm items, (b) some were new items, and (c) yet others were related to specific midterm items in that they required the same conceptual knowledge to answer but were phrased differently. Each of these three item types was further divided into two subsets. One subset of questions required students to recall facts or definitions from the course; the other subset required students to apply their knowledge to a new problem or setting. There were eight instances of each of these six subtypes on the final. The Appendix gives examples of original (midterm) and related (final exam) items in both the fact and application conditions. To summarize: six item subtypes occurred on

the final; they resulted from a 3 (item types: exact, related, new)  $\times$  2 (knowledge requirement to answer: fact-based, application via example) factorial design.

One interpretation of the testing effect literature makes a clear prediction about the exact items in both the factual and application modes: to the extent that a pattern match between the original and later items is helpful in retrieving the associated answer, we should see superior performance for exactly repeated items compared with new items. There is a question about the related items, however, and the literature on this has mixed messages to date. We will return to this issue in the General Discussion section. For practical reasons we certainly hope to see savings for the related test items—as educators we bank on transfer at least this far from the original learning. However, when presented in another cloak, it is not clear when, or even that, students will recognize the problem type; the changed cloak may render the required knowledge invisible.

#### Method

**Participants.** The participants were 214 undergraduate students at the sophomore level and above enrolled at the UH. There were 115 students in the standard exam section; of these, 43 were in the lab quiz sections and 72 were in the no-quiz lab sections. A total of 99 students were in the frequent exam lecture section; 67 of them were in the quiz sections and 32 were in the no-quiz lab sections.

**Materials and procedure.** In this study the frequent testing section was offered at 10:00 a.m. on Tuesdays and Thursdays, while the standard testing section took place at 11:30 a.m. on those same days. There were 28 class meetings across 15 weeks. The two midterms in the standard class took place in Weeks 6 and 12. Each contained 24 SA items. The eight in-course exams in the frequent class took place in Weeks 2, 4, 5, 6, 8, 10, 12, and 14. Each had six SA items. Overall, both the standard and the frequent classes were given the same 48 midterm items. Two forms of the test were used in each exam; in this study the two forms differed only in the order the questions were asked.

Just as in Study 3, students in the quiz sections were informed that the quizzes would be collected but they would not receive a grade on them. They were told that the quizzes were for practice and for their benefit. However, contrary to the method described in Study 3, in Study 4 the lab instructors announced quizzes on the course syllabus so students were aware of when they were going to take place. All quiz items were presented in SA format, thus matching the format used on the exams. The lab instructor provided feedback on the correct answers to each item immediately after each quiz. As in the previous study, the quizzes were administered in lab sections prior to Exams 1, 2, 3, 5, 7, and 8 for the frequent exam group and three of the six quizzes were given prior to Exam 1 for the standard group.

The final exam in the lecture sections contained 56 items. The first 48 were SA questions, eight from each of the six subtypes described above, that is, those resulting from a 3 (relationship to prior exam items: exact, related, new)  $\times$  2 (knowledge requirement to answer: fact-based, application via example) factorial design. (As noted earlier, the last 8 items were used to explore another hypothesis not discussed here.)



Results

The results in Figure 4 show the LS means for Testing Frequency  $\times$  Item Type. These results were produced by a repeated-measures ANCOVA. We used orthogonal contrasts to test (a) whether students perform better on repeated (exact and related) than on new items, (b) whether students performed better on exact items than on related ones, and (c) whether students perform better on fact-based items than on application items. As performed in Studies 1–3 we used an additional ANCOVA to directly test the effect of testing frequency and quiz status on final exam score.

As in previous analyses, student aptitude,  $F(1, 205) = 53.53$ ,  $p < .0001$ , was used as a covariate. Testing frequency,  $F(1, 205) = .68$ ,  $p = .41$ , and quiz status,  $F(1, 205) = 1.58$ ,  $p = .21$ , and the interaction of testing frequency and quiz status,  $F(1, 205) = .42$ ,  $p = .52$  were not significant. Results show a significant main effect of item type,  $F(5, 205) = 14.54$ ,  $p < .0001$ , and a significant interaction of testing frequency and item type,  $F(5, 205) = 3.91$ ,  $p = .002$ . Figure 4 shows performance on the final exam by testing frequency and item type. The interaction of item type and quiz status was not significant,  $F(5, 205) = 1.39$ ,  $p = .23$ .

Students performed reliably better on repeated (LS  $M = .55$ ) than on new items (LS  $M = .52$ ;  $d = .23$ ),  $F(1, 205) = 15.25$ ,  $p = .0001$ . However, students did not perform significantly better on exact items (LS  $M = .55$ ) than on related items (LS  $M = .56$ ;  $d = -.08$ ),  $F(1, 205) = 1.77$ ,  $p = .19$ . They did perform better on fact-based items (LS  $M = .56$ ) than on application items (LS  $M = .52$ ;  $d = .31$ ),  $F(1, 205) = 23.22$ ,  $p < .0001$ .

An ANCOVA on the proportion correct from the final exam found that the overall effect due to the frequency of in-class exams (frequent LS  $M = .55$ ; standard LS  $M = .55$ ;  $d = .04$ ) was not significant,  $F(1, 209) = 0.00$ ,  $p = .98$ . Student aptitude was significant,  $F(1, 209) = 48.74$ ,  $p < .0001$ . Further, in this study the effect on the final exam due to quiz status (no quiz LS  $M = .53$ , 6 quizzes LS  $M = .56$ ;  $d = .23$ ), did not reach statistical significance,  $F(1, 209) = 2.01$ ,  $p = .16$ ; though there was a hint of an interaction between testing frequency and quiz status,  $F(1, 209) = 3.02$ ,  $p = .08$ , as can be seen in Table 3. The LS means (standard

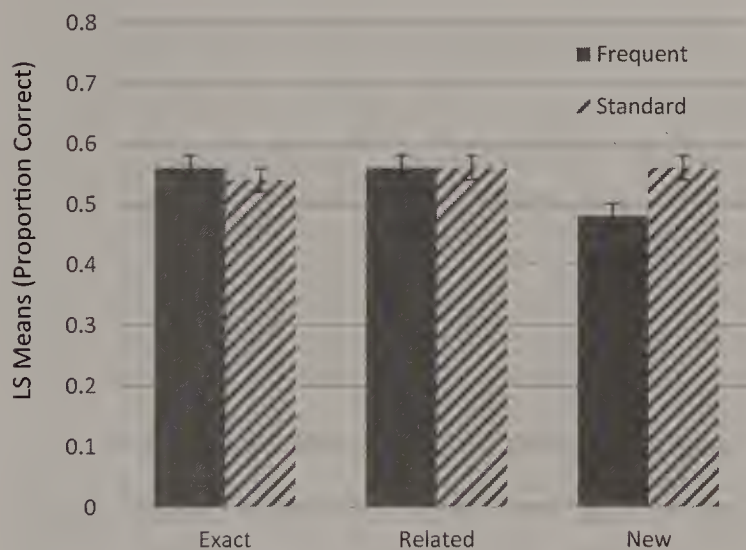


Figure 4. Least squares means (proportion correct) on the final exam in Study 4 as a function of testing frequency and item type. Error bars represent standard errors of the means.

Table 3  
LS Means (SD) on Cumulative Score by Testing Frequency and Quiz Condition for Study 4

Quiz condition	Frequent	Standard	Mean
Quiz	.54 (±.13)	.58 (±.14)	.56
No quiz	.55 (±.16)	.51 (±.15)	.53
Mean	.55	.55	

Note. Table shows the least squares (LS) means for cumulative score in each cell of the Testing Frequency  $\times$  Quiz interaction in Study 4.

deviations) for proportion correct on the final exam for each of the four combinations of lecture exams ( $S = \text{standard} = 2$ ;  $F = \text{frequent} = 8$ ) and number of lab quizzes (0, 6) are:  $S0 = .51 (.15)$ ;  $S6 = .58 (.14)$ ;  $F0 = .55 (.16)$ ; and  $F6 = .54 (.13)$ .

Discussion

Previous data, specifically the results of Studies 2 and 3 (where the largest differences between the frequent and standard classes were on final exam SA items), led us to use exams totally comprised of SA items. We expected even larger differences between the frequent and standard classes in Study 4, yet no difference between the two classes was observed. We will further discuss this finding below. Although main effects and interactions of testing frequency and quiz status were not significant, students in the standard class tended to benefit more from additional quizzes than students in the frequent class.

Though the effect of quizzing in Study 4 was not statistically significant, performance was in the predicted direction contrary to the results of Study 3. As previously mentioned, that may have been due to slight changes in our procedure. For example, in Study 4 the quizzes were announced in class and posted on the syllabus—but that was not the case in Study 3. It is also possible that we observed a larger effect of quizzing in Study 4 because all exam and quiz items were SA. Using this logic, students in the quizzing condition may have performed better because they had more practice answering recall items, which we know to be challenging for students.

With respect to the testing effect, the results due to item type in Study 4 replicated the findings in Studies 1–3 that students perform better on repeated than on novel items. Interestingly, we also found evidence for transfer effects in that there was no statistical difference between performance on exact and related items. That is, on the final exam students performed equally well on items worded exactly as previously seen and on items that asked a slightly different question (albeit using the same topical material) than was asked on the midterms. Not surprisingly, students performed better on fact-based items than on application items.

Overview of Results

Here we will provide a summary and overall analysis that captures the results across the four terms.

We calculated two primary measures of course performance: final exam score and cumulative score, the latter being the cumulative performance on all exams including the midterms and final exam. To calculate the cumulative score, each exam question



(including all questions on both the midterm exams and final exam) was given the same weight. Thus the cumulative score variable can be thought of as the overall percent correct throughout the semester. Although both of these outcome measures have some interest, the final exam score is generally the most straightforward one because in each semester the administration of the final exam was identical for all students, whereas the cumulative score includes data from the midterms where frequency of exam administration and length of the exams differed between the two classes—though as noted earlier the midterm items were identical across the classes in each semester. In addition, and importantly, the average time between presentation of the material and testing it on a midterm is shorter in the frequent than in the standard classes (though this was not true for the final exams).

The results presented here come from a total of 588 students for whom we had student aptitude scores and who completed the course. We omitted students who dropped the course, plus an additional 31 students who did not have an aptitude score. The latter were recent transfer students who had not made a GPA record at UH and also had no standard test scores in the UH records. (Though not reported here, an analysis that included those 31 students—where we imputed an average student aptitude score relative to other students in the study—showed slightly greater differences on the effects of interest.)

### Effects of Prior Testing on Subsequent Test Performance

A repeated measures ANCOVA (comparable to those reported above) was used to summarize the results of Studies 1–3 and examine the set of key questions motivating this work. Among them: (a) whether on the final exam students perform better on items that have been seen on previous exams, (b) whether performance on repeated items depends on whether the item form (MC or SA) was consistent across exams, (c) whether the effect of testing frequency depends on item type, and (d) whether the low-stakes quizzes affect final exam results. In addressing these issues, data from Study 4 were analyzed separately from Studies 1–3 because of differences in the study design, primarily the fact that exams in Study 4 were comprised totally of SA items.

As done in previous analyses, a proportion correct score for six types of items (MC-MC, MC-SA, SA-MC, SA-SA, new MC, and new SA) was calculated for each student. These outcome measures were regressed on predictors: testing frequency, quiz status, item type, student aptitude, and semester. We used specific contrasts to test hypotheses (a) and (b), and fixed effects results from the mixed model to test hypotheses (c) and (d).

Results revealed significant main effects of the covariates student aptitude,  $F(1, 328) = 112.56, p < .0001$ ; and semester,  $F(2, 328) = 3.56, p = .03$ . Item type,  $F(5, 328) = 75.48, p < .0001$ , and the interaction of testing frequency and item type,  $F(5, 328) = 4.34, p = .0008$ , were also significant (Table 4). The interaction of testing frequency and quiz status,  $F(1, 328) = .93, p = .34$ , was not significant; however, the interaction between quiz status and item type,  $F(5, 328) = 6.88, p < .0001$ , was significant.

We used orthogonal contrasts to test the additional hypotheses that:

(a) Students perform better on final exam items that previously appeared on a midterm exam (LS  $M = .71$ ) than on new items (LS

Table 4

*LS Means on Final Exam (Proportion Correct) for Testing Frequency by Item Type, Studies 1–3*

Test frequency	<i>n</i>	MC items			SA items		
		MC-MC	SA-MC	New MC	SA-SA	MC-SA	New SA
Frequent	193	.76	.78	.71	.65	.70	.60
Standard	142	.76	.80	.68	.59	.64	.57
Total	335	.76	.79	.70	.62	.67	.59

*Note.* The *ns* do not include the 214 students from Study 4; and 39 students from Study I who have missing data; MC-MC, SA-MC, and New MC were all multiple-choice (MC) items at the final exam; SA-SA, MC-SA, and New SA were all short-answer (SA) items at the final exam.

$M = .65; d = .32$ ). This repeated versus new item contrast was significant,  $F(1, 328) = 95.12, p < .0001$ .

(b) Superior performance would be observed for items that were repeated in the same format, for example, MC-MC and SA-SA (LS  $M = .69$ ) as opposed to being flipped, for example, SA-MC, MC-SA (LS  $M = .73; d = -.21$ ). That difference was also significant  $F(1, 328) = 24.42, p < .0001$ , though on average the flipped items performed better than the identical ones, contrary to expectations.

To examine hypothesis (c): whether the effect of testing frequency depends on item type, we used results from the same repeated measures described above to evaluate the main effects of testing frequency and item type and their interaction. Table 4 shows the LS means (proportion correct) on the final exams from Studies 1–3. More specifically, this table shows the mean proportion of items correct associated with each cell in the combination of testing frequency (frequent vs. standard) and item type (6 types of items); and the model tests the main effects of these two variables and their interaction. Results from this model show a significant main effect of item type,  $F(5, 328) = 75.48, p < .0001$ , and a significant interaction between testing frequency and item type,  $F(5, 328) = 4.34, p = .0008$ . The main effect of testing frequency,  $F(1, 328) = 3.11, p = .08$ , approached significance.

To examine hypothesis (d), whether the low-stakes quizzes affected final exam results, we used the fixed effects results from the same mixed linear model as was used to test hypotheses (a)–(c). The results show that while overall the quiz condition,  $F(1, 328) = .66, p = .42$ , failed to predict final exam performance, we did observe a significant interaction between quiz status and item type,  $F(5, 328) = 6.88, p < .0001$ . Table 5 shows the results (LS means) for the interaction of quiz and item type in Studies 1–3. Though not directly related to any specific hypothesis in this analysis, we observed that the effect of testing frequency did not depend on the level of the quiz variable,  $F(1, 328) = .93, p = .34$ .

### Frequency of Testing

To address questions about frequency of testing and the testing effect (do students perform better on previously tested items?), and the effect of low-stakes quizzes (do students perform better when given such quizzes?), we provide results from an ANCOVA (SAS 9.4) with four factors: student aptitude, testing frequency, quiz status, and semester. The semester variable was used to control for differences in the administration of each study, and of course the

Table 5  
LS Means on Final Exam (Proportion Correct) for Quiz by Item Type, Studies 1–3

Quiz condition	n	MC items			SA items		
		MC-MC	SA-MC	NEW MC	SA-SA	MC-SA	New SA
Quiz	91	.77	.80	.68	.61	.70	.62
No quiz	244 <sup>a</sup>	.75	.78	.72	.63	.64	.55
Total	335	.76	.79	.70	.62	.67	.59

<sup>a</sup> Includes participants from Studies 1 and 2, none of whom were given quizzes; MC-MC, SA-MC, and New MC were all multiple-choice (MC) items at the final exam; SA-SA, MC-SA, and New SA were all short-answer (SA) items at the final exam.

participants differed across semesters. We first fit a full model that includes all four factors mentioned above, and then we examine a reduced model that omits some nonsignificant factors to evaluate the effect of testing frequency in the most parsimonious context.

Final Exam Performance

**Full model.** Naturally, there was a large effect for student aptitude on final exam performance,  $F(1, 577) = 218.34, p < .0001$ , and also a significant main effect for Semester,  $F(3, 577) = 32.50, p < .0001$ . Table 6 shows both the raw means (proportion correct) for each condition and the LS mean for testing frequency in this model. (There were 64 final exam items in Studies 1–3 and 48 final exam items in Study 4.) The results from the full model on the final exam score indicate that testing frequency,  $F(1, 577) = 2.23, p = .14$ , was not significant. The interaction of semester and testing frequency was not significant,  $F(3, 577) = 1.78, p = .15$ . The main effect of quiz status was not significant (quiz LS  $M = .66$ ; no-quiz LS  $M = .63$ ),  $F(1, 577) = 1.65, p = .19$ , nor was the interaction between testing frequency and quiz status,  $F(1, 577) = .96, p = .33$ .

**Reduced model.** Evidence from the Type I sums of squares in the full model on final exam score suggested that testing frequency

exhibits a significant effect when controlling for student aptitude. For this reason, a reduced model was used to further examine the effect of testing frequency, while omitting nonsignificant factors (quiz status and the interaction of testing frequency and quiz status). This reduced model includes the factors: student aptitude, testing frequency, and semester, and the interaction of testing frequency and semester.

In the reduced model student aptitude,  $F(1,579) = 219.62, p < .0001$  and semester  $F(3, 579) = 5.29, p < .0001$  again show strong effects. We also observed a significant effect of testing frequency,  $F(1, 579) = 5.29, p = .02$ . The frequent class outperformed the standard class on the final exam in Studies 1–4 where we observed effect sizes (Cohen’s  $d$ ) of .07, .20, .28, and .04, respectively. Figure 5 shows the LS means for final exam score by condition in Studies 1-4. The interaction of semester and testing frequency,  $F(3, 579) = 1.37, p = .25$  was not significant.

Cumulative Exam Performance

**Full model.** When the outcome variable cumulative score (also shown in Table 6) is used, a parallel (full model) analysis again found that both student aptitude,  $F(1, 577) = 277.92, p < .0001$ , and semester,  $F(3, 577) = 14.75, p < .0001$ , had large effects. (Cumulatively, there were 104 test items in the Studies 1 and 4 and 112 in Studies 2 and 3.) For these data, even in the full model we observed a significant main effect of testing frequency,  $F(1, 577) = 6.83, p = .01$ . The interaction between testing frequency and semester was not significant,  $F(3, 577) = .77, p = .38$ . The difference in LS means for students in the quiz condition (.64) and the no -quiz condition (.63) was not statistically significant ( $d = .05$ ),  $F(1, 577) = .77, p = .38$  nor was the interaction of testing frequency and quiz status,  $F(1, 577) = 1.31, p = .35$ .

**Reduced model.** The reduced model on cumulative score (analogous to the reduced model on final exam performance) showed comparable results to the full model for the main effects of student aptitude and semester. However, the main effect of testing frequency,  $F(1, 579) = 13.15, p = .0003$ , was significantly stronger in the reduced model. As observed in the full model, the

Table 6  
Descriptive Statistics for Studies 1–4

Experiment	Condition (number of exams)	n	Final exam mean	Final exam LS mean	Final exam SD	Cohen’s $d$ on LS means final exam	Cumulative score mean	Cumulative score LS mean	Cumulative score SD	Cohen’s $d$ on LS means cumulative
Study 1	Frequent (4)	84	.66	.66	.15	.07	.67	.67	.13	.15
	Standard (2)	75	.64	.65			.64	.65		
Study 2	Frequent (8)	34	.73	.71	.15	.20	.71	.68	.14	.43
	Standard (2)	36	.66	.68			.63	.62		
Study 3	Frequent (8)	75	.69	.68	.18	.28	.68	.67	.16	.31
	Standard (2)	70	.62	.63			.61	.62		
Study 4	Frequent (8)	99	.55	.55	.13	.04	.56	.59	.15	.20
	Standard (2)	115	.53	.55			.52	.56		
Total	Frequent	292	.64	.65	.19	.16	.64	.65	.17	.24
	Standard	296	.59	.62			.58	.61		

*Note.* Final exam mean is the raw mean score (proportion correct) on the final exam; final exam least squares (LS) mean is the mean score on the final exam over a balanced population (e.g., controlling for differences in student aptitude); LS means estimates for both final exam and cumulative score are taken from the respective reduced models; cumulative score mean is the mean total score on all midterm exams and the final exam; Cohen’s  $d$  was calculated by the dividing the difference between the frequent and standard class LS mean by the pooled within-group standard deviation; the effect sizes for each study and the overall effect size for each measure are reported.



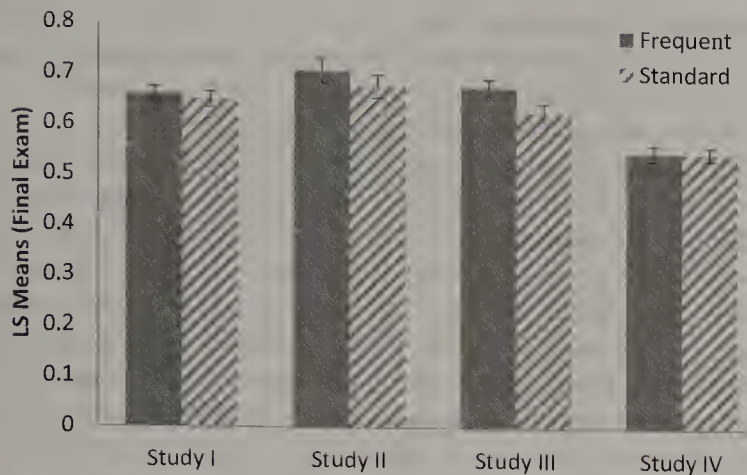


Figure 5. Least squares means (proportion correct) on the final exam in Studies 1–4 as a function of testing frequency. Error bars represent standard errors of the means.

frequent class consistently outperformed the standard class in Studies 1–4, where we estimate effect sizes ( $d$ ) of .15, .43, .31, and .20, respectively (see Table 6). Figure 6 shows the LS means for cumulative score by condition in Studies 1–4. The interaction of testing frequency and semester was not significant,  $F(3, 579) = .76, p = .52$ .

### General Discussion

While the methodology used in these four studies differed somewhat from semester to semester, there were fundamental conceptual similarities among the manipulations we carried out and considerable, though not total, consistency in the results as can be seen by again inspecting Tables 4 and 6.

We can summarize the overall results as follows:

1. More frequent testing generally, though not invariably, led to better final exam and overall course results.

This replicates a phenomenon that has been observed many times over the years, though not often in college classes over an entire semester, and rarely if ever before replicated in the fashion of this work which allows us to see a range of effect sizes along with variation of procedures within a common framework. In addition to the direct evidence, there is additional confirmation of the benefits due to frequent testing embedded in the results of the various methods used in our experiments. For example, we observed a larger effect size of frequent testing on the final exam in Study 2 ( $d = .20$ ) than in Study 1 ( $d = .07$ ). The main difference between these two studies was the increase in number of exams in the frequent class from four in Study 1 to eight in Study 2. Though we did not directly test the effect of four versus eight exams in any one study, this work provides evidence that (at least up to a point) more frequent testing improves performance on the final exam in a full semester course.

In Studies 2 and 3 the increase in performance of the frequent class over the standard class was not only reliable by usual standards, but was of meaningful size: 12%–17% on the total cumulative scores. In practice, that was an average improvement of better than 1/2 a letter grade in the current courses. To give a sense of the practical benefit of such improved performance, depending

upon the size of the class and, of course, on the grading scheme adopted, it could allow a consequential number of additional students to earn passing grades. That is a significant (in both senses) improvement at very little additional cost and effort. Note that the resources required to grade the exams, including time to do so, are quite similar across the two conditions. Thus, giving more frequent exams (e.g., 8 per semester) would be an inexpensive and easily adoptable and adaptable modification to many college courses.

Though we report a relatively consistent trend for the frequent class to outperform the standard class over four semesters, when we analyzed performance on the different items types we observed a significant interaction between testing frequency and item type in Studies 1–3. The data in Table 4 show that the frequent and standard classes did not differ on MC final exam items, but that the frequent class consistently outperformed the standard class on final exam SA items. This finding suggests that frequent testing may have large benefits for items that require recall, and a smaller effect on recognition items. However, that conclusion is considerably tempered by the results of Study 4 where the frequent class did not outperform the standard class on the final exam, and where the test items were in SA format. Perhaps the fact that the items were substantially more difficult than in the earlier studies (about 15% fewer correct answers, on average) played a role in this finding. With fewer correct answers, participants may have had less internal “good advice” to draw upon when attempting to answer final exam items. That is, less was learned from the exams themselves. This discrepancy in the overall findings for the frequency effect requires further unpacking in the future.

Furthermore, we also observed an interaction between quiz and item type. Students in the quiz condition showed superior performance on final exam SA items (see Table 5). As we defined them, testing frequency and quiz are primarily distinct in that testing frequency manipulates the distribution of exams (with the total number of exam items held constant), and quizzing provides addition retrieval practice. Thus students in the quiz condition (regardless of whether in the frequent or standard class) received the opportunity to answer 36 additional questions over the course

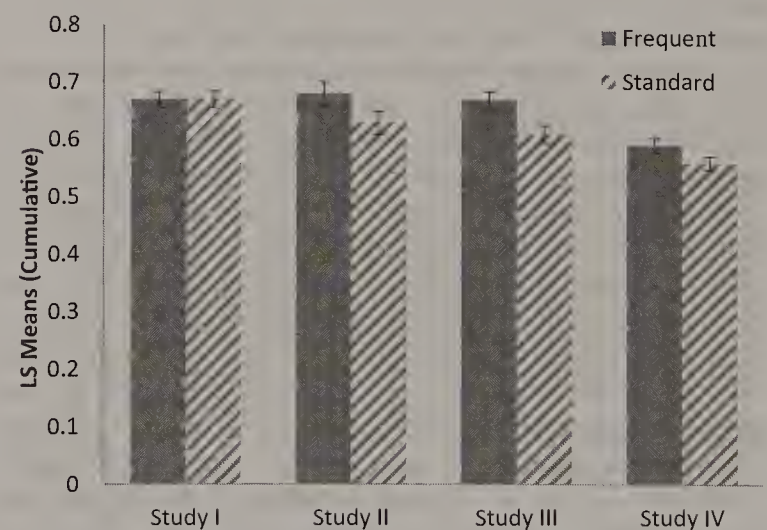


Figure 6. Least squares means (proportion correct) on cumulative score in Studies 1–4 as a function of testing frequency. Error bars represent standard errors of the means.



of the semester. This additional retrieval practice may be responsible for improved performance on SA (recall) items.

2. The repetition or testing effect is generally present and meaningful in these ecologically valid, semester-long studies.

In all four studies we found that students perform better on repeated items (those seen previously on midterm exams) than on items that had not been previously tested. This finding remained highly reliable even when in Study 4 we included related items in the repeated category. This finding replicates several findings that show testing or retrieval effects in practical educational settings. Our studies differ from most other research on the testing effect in that they each took place over an entire semester. These findings suggest that testing on material (even once) and receiving feedback can improve long-term retention.

3. The testing effect generally held, even when the format of the item changed (e.g., from MC on the first test to SA on the final, and vice versa).

The percentage of questions answered correctly on the final exam increased by about 8% for MC items, and about 5% for SA items in the testing effect conditions—that is, when the items had appeared on earlier exams. Again, these are meaningful increases in performance that make a difference in the grade distributions in large class sections.

It is also notable that participants in both the frequent and the standard groups improved when presented with a repeated item whether given in identical or flipped format. This finding is consistent with data from Butler (2010) and replicates observations made by Bjork et al. (2014), and by McDermott et al. (2014). Put another way, the testing effect transferred to items presented in an alternate format. Indeed, our students generally did somewhat better on the flipped items than on the identical ones. We consider that a surprising and likely important finding—one also remarked upon by McDermott et al. (2014)—though, as noted in the introduction, not everyone has observed it (e.g., Wooldridge et al., 2014). If it does exist, how would we account for the apparent superiority (or even equality) of flipped and related questions compared with identical ones asked on the final exam?

Bjork et al. (2014) “found that test-takers ability to answer related questions on a delayed test was only enhanced when the correct answers to such questions had been plausible incorrect alternative on the previous test . . .” (p. 169). They suggest that students who process those alternative answers learn distinctions that can help them answer subsequent, related questions. An inspection of Table 4 suggests some support for that idea: namely, an increased probability of getting an SA item correct on the final exam when preceded by a flipped MC item (.67) than when preceded by the same SA item (.62). However, there was no difference in probability of getting an MC item correct on the final exam whether preceded by an MC (.76) or an SA (.79) item. Admittedly, we may be observing a ceiling effect here; and our items do not really parallel those used by Bjork et al. (2014).

Stepping back, there is an extensive literature on transfer of training, one of the oldest topics in experimental psychology (e.g., Barnett & Ceci, 2002; Detterman & Sternberg, 1993; Harrison,

Shipstead, & Engle, 2015; Mayer & Wittrock, 1996; Singley & Anderson, 1989; Taatgen, 2013). Almost any explanation of transfer would predict that an identical test item should more readily be associated with the previously presented one than would a changed test item. All the surface cues are the same in the identical case. However, in our studies that result appears not to hold. What’s different? For one thing, in this work a great deal of time—up to 3 months—could pass between the original test and the final one. In addition, the intervening time is not just filled with normal everyday life going by. Importantly, we think, additional related and relevant material is presented to the participants during that time and they are building an extensive and interrelated knowledge base. Too, they attempt to retrieve or recall information from that knowledge base on subsequent exams and on other occasions (e.g., working homework problems on basic statistical concepts). Under those circumstances, maximally effective interrogation of the complex knowledge representation may not require pattern matching with the original form of the question.

To make an accurate prediction of transfer, we may need to consider both the length of time since initial presentation and, importantly, the structure of the resulting representation—one constructed from the intervening material as well as from the material tested on the early exams. Thus, there may be multiple effective cues that can access prior information, now a component of a more complex representational system. In addition, we need a measure of how well the new question overlaps with that changing representation. Thus, what appears as a “far” transfer item on the surface may in fact be a “near” one when we compare it to the latent representation of what the student knows (see, e.g., Taatgen, 2013, for a discussion of far transfer in the skill domain).

In sum, should this finding hold up, it again points to the need for better understanding of transfer, especially with long time delays partially filled with complex, interrelated materials. After all, our goal as educators is to present and test information such that it leads to a greater likelihood of being available to aid both question answering and problem solving at much later times.

4. We did not observe an overall benefit from taking the low-stakes quizzes.

The failure to see benefits of low-stakes quizzes could be due to several factors. Our results showed no difference (or even a negative effect) between performance in the quiz and no-quiz condition in Study 3. However, in Study 4 we showed a slight advantage to taking quizzes. As previously mentioned, advance announcement of the quizzes and the testing format used in Study 4 could have influenced the results. Because no grade depended on individual quiz performance there may also be motivational factors that modulate the effect of low-stakes quizzes. An extensive and wide-range of research on motivation and performance shows evidence for individual differences in the origin of motivation to perform in educational tasks (e.g., Wolters, Denton, York, & Francis, 2014). Studies in the motivation literature also show how incentives can play a large role in performance (e.g., Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011). If motivation and incentives can modulate performance on exams, it is intuitive that these factors play a role in the effectiveness of educational training conditions such as those employed in the present work. So here, though not only here, we endorse the



recommendation that further research is necessary for both theoretical reasons and to explore the limits of practical applications of this work.

One methodological point is worth revisiting: The inability to randomly assign participants to conditions poses a potential limitation to our conclusions. We dealt with this issue by obtaining covariates for student aptitude (GPA, SAT, and ACT scores) and including them in our analyses, and by alternating the time (10:00 a.m. or 11:30 a.m.) of the testing frequency treatment across semesters. Furthermore, we acknowledge the limitation that testing frequency systematically varied with the classroom—that is, effects of classroom clustering (classroom environment) could have influenced our results. Given that we used one instructor, we are unable to estimate effects due to class cluster.

In summary, the findings in Studies 1–4 suggest that frequent testing and making use of the testing effect can improve performance on a comprehensive final exam (and the cumulative course performance) in “live” classroom settings over an entire semester. It appears that typically, but not inevitably (perhaps when item difficulty was very high), the effect of frequent testing is greatest on difficult (e.g., SA) items that require recall rather than recognition. For the most part, these appear to be inexpensive and readily scalable findings. We also found that the benefit of prior testing was not limited to items that are exactly repeated, but appeared to generalize to new questions that tapped the same information in order to answer them. In addition, we observed small to nonexistent effects of no-stakes quizzes, but certainly feel that this matter requires further exploration for the reasons noted above.

Finally, we acknowledge that there is much left to do before we understand the differences in effect sizes that we and others have observed, and to explore further the transfer of training issues that arise both in laboratory studies and larger scale investigations such as reported here.

## References

- Bahrack, H. P., Bahrack, L. E., Bahrack, A. S., & Bahrack, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, 4, 316–321. <http://dx.doi.org/10.1111/j.1467-9280.1993.tb00571.x>
- Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research*, 85, 89–99. <http://dx.doi.org/10.1080/00220671.1991.10702818>
- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128, 612–637. <http://dx.doi.org/10.1037/0033-2909.128.4.612>
- Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory & Cognition*, 3, 165–170. <http://dx.doi.org/10.1016/j.jarmac.2014.03.002>
- Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 1118–1133. <http://dx.doi.org/10.1037/a0019902>
- Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, 24, 369–378. <http://dx.doi.org/10.1007/s10648-012-9205-z>
- Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, 23, 760–771. <http://dx.doi.org/10.1002/acp.1507>
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132, 354–380. <http://dx.doi.org/10.1037/0033-2909.132.3.354>
- Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridge of optimal retention. *Psychological Science*, 19, 1095–1102. <http://dx.doi.org/10.1111/j.1467-9280.2008.02209.x>
- Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The European Journal of Cognitive Psychology, 21, 919–940. <http://dx.doi.org/10.1080/09541440802413505>
- Delaney, P. F., Verkoeijen, P. L., & Spigel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross & B. H. Ross (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 53, pp. 63–147). San Diego, CA: Elsevier Academic Press. [http://dx.doi.org/10.1016/S0079-7421\(10\)53003-2](http://dx.doi.org/10.1016/S0079-7421(10)53003-2)
- Detterman, D. K., & Sternberg, R. J. (1993). *Transfer on trial: Intelligence, cognition, and instruction*. Westport, CT: Ablex.
- Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, 84, 795–805. <http://dx.doi.org/10.1037/0021-9010.84.5.795>
- Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 108, 7716–7720. <http://dx.doi.org/10.1073/pnas.1018601108>
- Dunlosky, J., & Rawson, K. A. (2012). Despite their promise, there's still a lot to learn about techniques that support durable learning. *Journal of Applied Research in Memory & Cognition*, 1, 254–256. <http://dx.doi.org/10.1016/j.jarmac.2012.10.003>
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14, 4–58. <http://dx.doi.org/10.1177/1529100612453266>
- Gaynor, J., & Millham, J. (1976). Student performance and evaluation under variant teaching and testing methods in a large college course. *Journal of Educational Psychology*, 68, 312–317. <http://dx.doi.org/10.1037/0022-0663.68.3.312>
- Glover, J. A. (1989). The ‘testing’ phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, 81, 392–399. <http://dx.doi.org/10.1037/0022-0663.81.3.392>
- Harrison, T. L., Shipstead, Z., & Engle, R. W. (2015). Taxonomy of transfer to cognitive abilities: The case of working memory training. In D. S. Lindsay, A. P. Yonelinas, H. I. Roediger, D. S. Lindsay, A. P. Yonelinas, & H. I. Roediger (Eds.), *Remembering: Attributions, processes, and control in human memory: Essays in honor of Larry Jacoby* (pp. 307–321). New York, NY: Psychology Press.
- Kang, S. H., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, 18, 998–1005. <http://dx.doi.org/10.3758/s13423-011-0113-x>
- Keys, N. (1934). The influence on learning and retention of weekly as opposed to monthly tests. *Journal of Educational Psychology*, 25, 427–436. <http://dx.doi.org/10.1037/h0074468>
- Kika, F. M., McLaughlin, T. F., & Dixon, J. (1992). Effects of frequent testing of secondary algebra students. *The Journal of Educational Research*, 85, 159–162. <http://dx.doi.org/10.1080/00220671.1992.9944432>

- Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education*, 43, 21–27. <http://dx.doi.org/10.1111/j.1365-2923.2008.03245.x>
- Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, 29, 210–212. [http://dx.doi.org/10.1207/S15328023TOP2903\\_06](http://dx.doi.org/10.1207/S15328023TOP2903_06)
- Mawhinney, V. T., Bostow, D. E., Laws, D. R., Blumenfeld, G. J., & Hopkins, B. L. (1971). A comparison of students studying-behavior produced by daily, weekly, and three-week testing schedules. *Journal of Applied Behavior Analysis*, 4, 257–264. <http://dx.doi.org/10.1901/jaba.1971.4-257>
- Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. C. Berliner, R. C. Calfee, D. C. Berliner, & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 47–62). New York, NY: Macmillan Reference USA.
- McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, 103, 399–414. <http://dx.doi.org/10.1037/a0021782>
- McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, 14, 200–206. <http://dx.doi.org/10.3758/BF03194052>
- McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, 27, 360–372. <http://dx.doi.org/10.1002/acp.2914>
- McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., III, & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, 20, 3–21. <http://dx.doi.org/10.1037/xap0000004>
- Myers, G. C. (1914). Recall in relation to retention. *Journal of Educational Psychology*, 5, 119–130. <http://dx.doi.org/10.1037/h0075769>
- Nguyen, K., & McDaniel, M. A. (2015). Using quizzing to assist student learning in the classroom: The good, the bad, and the ugly. *Teaching of Psychology*, 42, 87–92. <http://dx.doi.org/10.1177/0098628314562685>
- Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning: A practice guide* (Report No. NCER 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from [http://ies.ed.gov/ncee/wwc/pdf/practice\\_guides/20072004.pdf](http://ies.ed.gov/ncee/wwc/pdf/practice_guides/20072004.pdf) <http://dx.doi.org/10.1037/e607972011-001>
- Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily online testing in large classes: Boosting college performance while reducing achievement gaps. *PLoS ONE*, 8, e79774. <http://dx.doi.org/10.1371/journal.pone.0079774>
- Rawson, K. A., & Dunlosky, J. (2013). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology: General*, 142, 1113–1129. <http://dx.doi.org/10.1037/a0030498>
- Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, 17, 382–395. <http://dx.doi.org/10.1037/a0026252>
- Roediger, H. L., III, & Karpicke, J. D. (2006a). Test-enhanced learning. *Psychological Science*, 17, 249–255. <http://dx.doi.org/10.1111/j.1467-9280.2006.01693.x>
- Roediger, H. L., III, & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181–210. <http://dx.doi.org/10.1111/j.1745-6916.2006.00012.x>
- Ross, C. C., & Henry, L. K. (1939). The relation between frequency of testing and progress in learning psychology. *Journal of Educational Psychology*, 30, 604–611. <http://dx.doi.org/10.1037/h0055717>
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140, 1432–1463. <http://dx.doi.org/10.1037/a0037559>
- Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.
- Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, 120, 439–471. <http://dx.doi.org/10.1037/a0033138>
- Wolters, C. A., Denton, C. A., York, M. J., & Francis, D. J. (2014). Adolescents' motivation for reading: Group differences and relation to standardized achievement. *Reading and Writing*, 27, 503–533. <http://dx.doi.org/10.1007/s11145-013-9454-3>
- Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory & Cognition*, 3, 214–221. <http://dx.doi.org/10.1016/j.jarmac.2014.07.001>



## Appendix

### Example of Related Items From Study 4

#### Fact-Based (Definition) Items

Midterm: if another variable systematically co-varies with the independent variable, then we likely have a \_\_\_\_\_

Final: A confound likely exists when an extraneous variable systematically co-varies with the \_\_\_\_\_

Midterm: Inferential statistics allow us to draw conclusions about the \_\_\_\_\_ on the basis of data from a \_\_\_\_\_

Final: When we draw a conclusion about a population on the basis of a sample, we are using \_\_\_\_\_ statistics.

#### Application Items

Midterm: Mr. Thinblood was born in Houston and moved to New York. He found himself feeling tired in the winter—more tired than he remembered being in Houston during the winter. He wondered whether cold weather makes people tired. How would we best ask this question scientifically (in standard form)? \_\_\_\_\_

Final: Mr. C. D. Cloud believes that whenever he washes his car it greatly increases the chances of rain. If he turned his belief into a scientific question, how would he express it (in standard form)? \_\_\_\_\_

Midterm: A person is at the 84th percentile on a standardized test. Given that the sample mean is 60 and  $SD = 9$ , what is this person's raw score? \_\_\_\_\_

Final: On an exam the mean is 50 and the  $SD = 5$ . You ended up at the 16th percentile on this exam. What score did you make? \_\_\_\_\_

Received May 25, 2016

Revision received January 30, 2017

Accepted February 1, 2017 ■

#### E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <https://my.apa.org/portal/alerts/> and you will be notified by e-mail when issues of interest to you become available!

# Learning-Related Cognitive Self-Regulation Measures for Prekindergarten Children: A Comparative Evaluation of the Educational Relevance of Selected Measures

Mark W. Lipsey  
Vanderbilt University

Kimberly Turner Nesbitt  
University of New Hampshire

Dale C. Farran  
Vanderbilt University

Nianbo Dong  
University of Missouri-Columbia

Mary Wagner Fuhs  
University of Dayton

Sandra Jo Wilson  
Abt Associates

Many cognitive self-regulation (CSR) measures are related to the academic achievement of prekindergarten children and are thus of potential interest for school readiness screening and as outcome variables in intervention research aimed at improving those skills in order to facilitate learning. The objective of this study was to identify learning-related CSR measures especially suitable for such purposes by comparing the performance of promising candidates on criteria designed to assess their educational relevance for pre-K settings. A diverse set of 12 easily administered measures was selected from among those represented in research on attention, effortful control, and executive function, and applied to a large sample of pre-K children. Those measures were then compared on their ability to predict achievement and achievement gain, responsiveness to developmental change, and concurrence with teacher ratings of CSR-related classroom behavior. Four measures performed well on all those criteria: Peg Tapping, Head-Toes-Knees-Shoulders, the Kansas Reflection-Impulsivity Scale for Preschoolers, and Copy Design. Two others, Dimensional Change Card Sort and Backwards Digit Span, performed well on most of the criteria. Cross-validation with a new sample of children confirmed the initial evaluation of these measures and provided estimates of test–retest reliability.

## *Educational Impact and Implications Statement*

The ability of prekindergarten children to regulate such cognitive functions as attention and task persistence is related to their learning and academic achievement. This study identified measures of such learning-related cognitive self-regulation especially suitable for screening pre-k children for school readiness and as outcome measures for interventions aimed at improving those skills.

**Keywords:** cognitive self-regulation, executive function, school readiness, measurement

The ability of young children to exert control over their cognition and behaviors within educational contexts has been variously labeled approaches to learning (Davoudzadeh, McTernan, &

Grimm, 2015; Zimmerman, 1990), learning dispositions (Katz, 1993, 2002), and work-related skills (Cooper & Farran, 1988, 1991; Schmitt, Pratt, & McClelland, 2014). However labeled,

This article was published Online First April 6, 2017.

Mark W. Lipsey, Peabody Research Institute, Vanderbilt University; Kimberly Turner Nesbitt, Department of Human Development and Family Studies, University of New Hampshire; Dale C. Farran, Peabody Research Institute, Vanderbilt University; Nianbo Dong, Department of Educational, School, and Counseling Psychology, University of Missouri-Columbia; Mary Wagner Fuhs, Department of Psychology, University of Dayton; Sandra Jo Wilson, Social and Economic Policy Division, Abt Associates.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080079 to

Vanderbilt University. Kimberly Turner Nesbitt and Mary Wagner Fuhs were supported by an Institute of Education Sciences Postdoctoral Training Grant (R305B100016). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Special thanks to Deanna Meador and the Peabody Research Institute research team for their very capable management of the data collection for this project.

Correspondence concerning this article should be addressed to Mark W. Lipsey, Peabody Research Institute, Vanderbilt University, Nashville, TN 37203. E-mail: [mark.lipsey@vanderbilt.edu](mailto:mark.lipsey@vanderbilt.edu)



ample research has demonstrated that children's ability to focus on classroom tasks, persist despite difficulty, and engage in learning activities are positively related to academic achievement (Duncan et al., 2007; Li-Grining, Votruba-Drzal, Maldonado-Carreño, & Haas, 2010; McClelland, Morrison, & Holmes, 2000; Morgan, Farkas, & Wu, 2011). The constellation of skills that support this behavior can be referred to broadly as cognitive self-regulation.

Research on cognitive self-regulation (CSR) has been conducted within various conceptual frameworks including attention, executive function, and effortful control. Attentional functions such as conscious detection and sustained focus on a target stimulus are foundational aspects of one's ability to control thoughts and behaviors (Posner & Rothbart, 2000; Rothbart & Ahadi, 1994). Executive function, in turn, is generally defined as a set of cognitive abilities that aid in the completion of goal-directed actions (Hughes & Ensor, 2011; Miyake et al., 2000). These abilities include adapting or shifting actions to changing situational demands (Zelazo, Frye, & Rapus, 1996), active maintenance and manipulation of information in working memory (Baddeley & Hitch, 1974), and inhibition of inappropriate but prepotent responses (Diamond, 1990). Related to inhibitory control is the construct of effortful control, which involves volitional behavioral regulation related to aspects of temperament (Kochanska, Murray, & Harlan, 2000).

A number of assessments of CSR-related constructs suitable for administration directly to pre-K age children have been developed within these research contexts, and many of them have been shown to be related to concurrent or future academic achievement (Allan & Lonigan, 2011; Blair & Razza, 2007; Gathercole, Brown, & Pickering, 2003; Jacob & Parkinson, 2015; Lan, Legare, Ponitz, Li, & Morrison, 2011) and achievement gains during the pre-K and kindergarten years (Fuhs, Nesbitt, Farran, & Dong, 2014; Matthews, Ponitz, & Morrison, 2009; McClelland et al., 2007; Ponitz, McClelland, Matthews, & Morrison, 2009; Welsh, Nix, Blair, Bierman, & Nelson, 2010). Indeed, evidence indicates that cognitive self-regulation measures are among the strongest predictors of achievement after prior measures of achievement itself (Duncan et al., 2007).

Aside from whatever theoretical insights derive from this research, the relations of such measures to the academic achievement of pre-K children has particular importance in educational contexts. Most immediately, the measures most strongly related to achievement might be used in assessments of school readiness to identify children whose CSR skills may not be sufficient to support effective engagement in the learning opportunities in kindergarten and beyond and thus need help enhancing those skills. Those measures, in turn, would be appropriate targets as outcome variables in intervention research aimed at finding ways those skills can be improved to better facilitate learning in classroom environments.

But, which of the many measures of CSR that can be used with pre-K aged children are especially suitable for these purposes? Most prior studies reporting relations with achievement have focused on only a few measures and typically did not have comparison of those relations as their main objective. Moreover, different studies have used different achievement measures and different samples, features that could themselves influence the magnitude of the relations, thus making it difficult to compare the performance of measures used in different studies. And no study has systematically assessed CSR measures with regard to the multiple attributes

that would make them most educationally relevant to pre-K students.

The purpose of the study reported here is to make just such a comparative assessment for a group of candidate measures selected to represent a range of CSR skills while also being easily administered to young children. The aim of this assessment is to identify CSR measures with clear educational relevance for pre-K children; that is, measures that perform especially well when the interplay between CSR and achievement is of interest in pre-K classroom settings. The results, in turn, are intended to provide guidance to pre-K researchers and practitioners seeking measures of CSR for screening or research applications that have sound measurement properties and demonstrated relations to learning and CSR-related classroom behavior.

### Criteria for Evaluating CSR Measures

A comparative assessment focused on the educational relevance of CSR measures for pre-K students first requires decisions about the basis for selecting candidate measures and specification of appropriate criteria with which to assess them. To identify promising candidate measures, we did not attempt to apply strict selection criteria but used the informed judgment of our research team to pick measures that represented a range of CSR skills and tasks (described in more detail below) and to favor measures more widely known and used in early childhood research. Further, with practical application and broad utility in mind, we considered only measures that could be easily administered in school settings by school personnel or researchers with limited resources; that is, those that could be completed in a relatively brief period without specialized equipment or online Internet connections. A similar assessment of computer-based CSR measures would be informative, but for this study we chose to focus on readily accessible measures so the findings would be as broadly useful as possible.

To assess the relative performance of the selected CSR measures, we identified a set of attributes we judged to be indicative of their educational relevance for use in pre-K contexts. The most important of these, of course, involved the relation of the measures to academic achievement. Three types of relations were differentiated. First, we examined correlations between the CSR measures administered at the beginning of the pre-K year and later achievement. With our focus on CSR skills related to learning, the most educationally relevant measures are those most directly predictive of achievement. Less predictive measures, by definition, are less closely associated with whatever influence CSR skills have on achievement.

Second, we compared the candidate measures on their ability to predict the gains in achievement made during subsequent periods. Children with better initial CSR skills may show higher subsequent achievement, but that does not necessarily mean they also gain more during that period. They are likely to have higher achievement levels to begin with and may simply maintain their relative position. If we expect children with better CSR skills to be better able to engage in the learning opportunities presented in pre-K classrooms, they should show greater gains in achievement over the pre-K year and, similarly, over later school years. The most educationally relevant CSR measures, therefore, should be those that show the strongest relations to subsequent achievement gains.



Third, we compared the candidate CSR measures on an even more specific kind of relation with achievement. Pre-K experience is not only expected to affect achievement but may also affect CSR skills themselves such that those skills will improve during the course of the school year. Indeed, learning to pay attention, stay on task, change tasks when asked, and other such CSR-related behaviors are part of the school readiness objectives of many pre-K programs. If CSR skills are related to gains in achievement, then gains in CSR skills should, in turn, be related to further gains in achievement. We therefore compared the candidate measures on the extent to which the CSR gains observed over the pre-K year were related to achievement gains. Those relations are especially informative about the potential of the different CSR measures as outcomes for research on pre-K interventions aimed at enhancing learning-related CSR skills. Such interventions would naturally want to target CSR skills for which there was some assurance that gains on those skills were associated with learning gains.

The CSR measures most on target for use in pre-K settings when their implications for learning and achievement are of primary interest should be those that show the strongest relations of these different kinds. That is, we would expect children with better learning-related CSR skills not only to have higher achievement, but to show greater achievement gains over time, and if those CSR skills improve, to show correspondingly larger gains in achievement. The more relevant measures of these learning-related CSR skills, therefore, are those that best demonstrate these relations.

We then brought two additional perspectives to the assessment of the educational relevance of the candidate CSR measures. For one of these, we considered the extent to which the measures were responsive to developmental change, that is, showed nontrivial increases as CSR skills improved through maturation and whatever facilitation occurred in school classrooms. CSR measures that show no or limited increases during pre-K and subsequent early grades are thus relatively insensitive to the gains young children are known to make during those periods. Measures that are more sensitive to change will, by their very nature, perform better for assessing change and distinguishing children whose CSR skills differ.

Finally, we considered the relation between the candidate measures and teacher ratings of the CSR-related learning skills they are able to observe in the classroom, including persistence, independence, organization, and participation. Teacher ratings of such learning skills have been found to be predictive of later academic achievement (Bodovski & Farkas, 2007; Davoudzadeh et al., 2015; Schmitt et al., 2014) and reflect how CSR skills are manifest in children's classroom behavior. However, these ratings show distinct differences from the results of direct assessments of children's CSR skills (Fuhs, Farran, & Nesbitt, 2015; Matthews et al., 2009; Schmitt et al., 2014), and thus cannot be assumed to be equivalent measures of the underlying CSR skills of interest. Nonetheless, the candidate CSR measures with the greatest educational relevance in pre-K settings should also show close relations to teacher ratings of the learning skills those teachers observe in the classroom. Such relations help establish the ecological validity of the measures for use in pre-K contexts as well as giving them credibility with teachers who may use them.

To conduct a comparative assessment of the performance of direct assessments of CSR skills on these attributes, we selected a range of candidate measures as described in more detail below and

administered them to a large sample of children at the beginning and end of the pre-K year and again at the end of kindergarten. We then used those data to assess each measure for its ability to predict achievement and achievement gain, responsiveness to change over time, and correlation with teacher ratings. The best performing measures identified in those analyses were then administered to a new sample of children before and after the pre-K year to allow cross-validation of the findings from the initial sample and support collection of test-retest reliability data. The procedural details and results are described in the sections that follow.

## Method

To identify candidate measures, we first reviewed the literature on executive function, effortful control, attention, and self-regulation in an attempt to delineate the range of skills likely to be relevant to learning-related CSR. The skill domains distinguished for this purpose were

1. Sustained attention—attending to and sustaining focus on a task.
2. Attention shifting—shifting focus within or between tasks as situations demand.
3. Working memory—active maintenance and manipulation of information in memory.
4. Inhibitory control—volitional inhibition of a prepotent response in order to complete a task.
5. Effortful control—suppression of impulsive or premature responses when required by a task.

We then reviewed a wide range of CSR-related measures that have appeared in research with pre-K age children (a list of those measures is in Appendix A). We categorized each according to the skill domain that seemed most central to accomplishing the tasks the measure presented, relying heavily on the description of the measure in the associated literature. Of course, none of these are pure measures of the skills indicated by the labels we applied to the respective domains; they all tap into multiple overlapping skills. But sorting them this way and selecting at least one measure from each category ensured that we would end up with a diverse set that collectively should span the full range of CSR skill domains identified in research on this topic. When making these selections, we prioritized measures previously shown to be related to academic achievement and those we judged to be most practical for administration in classroom settings without the need for computer support or specialized equipment. Through this process, we identified 10 candidate measures that yield 12 indices of CSR (two measures assess both accuracy and reaction time [RT]), which are described below.

## Sustained Attention

For assessment tasks requiring the capacity to maintain focus and attention, we chose Copy Design (Davie, Butler, & Goldstein, 1972; Osborn, Butler, & Morris, 1984) and the Kansas Reflection-Impulsivity Scale for Preschoolers (KRISP; Wright,



1971). For Copy Design, children copy eight geometric designs of increasing difficulty and, for each, the quality of the best attempt is scored 0 or 1 by defined criteria with total scores ranging from 0 to 8. Cronbach alphas for this measure were .79 in the data we collected on our sample (described below) at the beginning of the pre-K year and .75 in the data collected at the end of the year.

The KRISP presents children with a series of drawings for which they must identify the duplicate of a target picture from 4–6 other pictures, all but one different in minor ways. Each of 12 trials is scored for number of errors and RT to selection of the first drawing. Accuracy is scored as the number of errors subtracted from the total errors possible (36). RT is scored as the difference between the mean for the 5 hardest and 7 easiest trials divided by the mean for the hardest ones, thus indexing how much the child slowed down to reflect on the harder items. Cronbach alphas for accuracy were .66 at the beginning of pre-K and .63 at the end.

### Attention Shifting

For measures requiring the ability to shift focus from one task to another, we selected the *Dimensional Change Card Sort* (DCCS; Zelazo, 2006). Children sort a set of cards according to one dimension (color), and then according to a different dimension (shape). If they are largely successful with that switch, they are given similar cards with a black border around some and asked to sort by color if the card has a border and by shape if not. Children receive a score of 0 if they do not pass the initial color sort, 1 if they pass the color but not the shape sort, 2 if they pass the shape sort, and 3 if they also pass the border sort. Cronbach alphas for color sorting were .81 for data at the beginning of pre-K and .78 at the end; for shape sorting, .96 at the beginning of pre-K and .92 at the end. Too few children were able to complete the border task to allow alpha values to be computed.

### Working Memory

For assessment tasks that require the ability to temporarily store and manage information, we selected Operation Span (Blair & Willoughby, 2006f) and Backwards Digit Span (Davis & Pratt, 1995). For Operation Span, children are shown pictures of houses with animals and colors and asked to name them, then recall the animal in each house on a second display of empty houses. Six trials with two, three, or four items to remember are scored 0 for incorrect and 1 for a correct response, with the sum as the final score (range 0 to 18). Cronbach alphas were .77 at the beginning of pre-K year and .64 at the end.

Backwards Digit Span (Davis & Pratt, 1995) asks children to remember, then reverse a series of numbers presented orally; for example, given 1, 3, the child is to respond 3, 1. Across six trials with increasing numbers of digits, each number recalled correctly in backward sequence is scored 1 with the final score as the sum of digits correctly recalled. In the pre-K year, too few children were able to complete a sufficient number of items for Cronbach's alpha to be computed.

**Inhibitory Control.** For measures that require the ability to suppress a prepotent response in order to complete a task, we selected Head-Toes-Knees-Shoulders (HTKS; Ponitz et al., 2009), Peg Tapping (Diamond & Taylor, 1996), and Spatial Conflict

(Blair & Willoughby, 2006e). HTKS asks a child to respond to oral prompts of “touch your head” and “touch your toes” by doing the opposite for 10 trials. If responses on five or more are correct, two new prompts are added for another 10 trials. Each trial is scored 0 for an incorrect response, 1 for an incorrect motion that was corrected, and 2 for a correct response with the sum across all items as the final score (range 0 to 40). Cronbach alphas for the first 10 trials were .96 at the beginning of pre-K and the same at the end; for the second 10, they were .85 at the beginning and .88 at the end.

The Peg Tapping task asks children to tap once when the examiner taps twice and twice when the examiner taps once (Diamond & Taylor, 1996). Children largely successful in practice trials then have 16 test trials scored 0 for incorrect and 1 for correct responses. Final scores range from –1 to 16, with –1 assigned if the child does not reach criterion in the practice trials. Cronbach alphas were .87 in data at the beginning of pre-K year and .88 at the end.

The Spatial Conflict task (Blair & Willoughby, 2006e) was a paper adaptation of the computer-based version (Gerardi-Caulton, 2000). Children are given a card with one button on the right-hand side and one on the left, and shown a series of arrows that point either left or right. They are asked to touch the button on the side the arrow points to using their right hand for the button on the right and their left hand for the one on the left. A series of congruent trials (arrow on the same side of the page it points to), is followed by 16 mixed congruent and incongruent trials scored 0 for the incorrect button, 1 for the correct button with the wrong hand, and 2 for the correct button with the correct hand, with the total score ranging from 0 to 32. Cronbach alphas were .82 for data from the beginning of pre-K and .77 at the end.

### Effortful Control

For assessment tasks that require the ability to suppress impulsive or premature responses, we selected the Whisper and Turtle-Rabbit tasks (Kochanska, Murray, Jacques, Koenig, & Vandegeest, 1996). In the Whisper task children are shown pictures of 12 cartoon characters and asked to whisper their names. The cartoon characters vary in familiarity, providing the opportunity for the child to act impulsively (shout) when a very recognizable one comes up. Each trial is scored 0 for a shout, 1 for a normal voice, 2 for no response, and 3 for a whisper (range 0 to 36). Cronbach alphas were .96 at the beginning of pre-K and .95 at the end.

The Turtle-Rabbit task (Kochanska et al., 1996) presents children with a drawing of a curved path with five bends and they are asked to move toy figures along the path without straying. After baseline trials with neutral figures, they are given two trials with a rabbit they are told is fast, and two with a turtle they are told is slow. Each curve is scored 0 if bypassed, 1 if the figure is above the mat but follows the general curvature, and 2 if the figure stays on the mat and within the path. Time to complete each trial is also recorded. Accuracy is scored as the total for all curves and trials (range 0 to 60). Reaction time is scored as the difference between the mean times for the turtle and rabbit trials. Cronbach alphas for accuracy were .99 for both rabbit and turtle at the beginning of pre-K, and .92 for rabbit and .89 for turtle at the end of pre-K.



## Teacher Ratings of Cognitive Self-Regulation

Teacher rating scales for children's behaviors in the classroom were selected to mirror as much as possible the aspects of CSR identified in our initial literature review and assessed in the candidate direct child measures. The following subscales were combined in a single rating form.

**Persistence.** The Persistence subscale of the Temperament Assessment Battery for Children (TABC; Martin, 1988) assesses each child's ability to sustain attention. The eight items on this subscale are rated on a 1 (*hardly ever*) to 7 (*almost always*) scale and include such behaviors as "child can continue at the same activity for an hour" and "if child's activity is interrupted, he/she tries to go back to the activity." Cronbach alphas for this subscale were .75 at the beginning of pre-K and .74 at the end.

**Distractibility.** The Distractibility subscale of the TABC assesses the ability to ignore distractions. The eight items on this subscale are rated as described above and cover such behaviors as "Child is easily drawn away from his/her work by noises . . . etc." and "If other children are talking or making noise while teacher is explaining a lesson, this child remains attentive to the teacher." Cronbach alphas were .89 at the beginning of pre-K and .90 at the end.

**Impulsivity.** This was assessed with the Impulsivity subscale of the Children's Behavioral Questionnaire (CBQ; Rothbart, Ahadi, Hershey, & Fisher, 2001). CBQ items are rated from 1 (*extremely untrue of student*) to 7 (*extremely true*). The 13 items cover such behavior as "sometimes interrupts others when they are speaking" and "usually stops and thinks things over before deciding to do something." Cronbach alphas were .87 at the beginning of pre-K and .88 at the end.

**Attention shifting.** The CBQ Attention Shifting subscale was used for this dimension. Twelve items are also rated as described above and include such behaviors as "needs to complete one activity before being asked to start on another one" and "can easily shift from one activity to another." Cronbach alphas were .87 at the beginning of pre-K and .89 at the end.

**Work-related skills.** A scale that spanned a variety of children's CSR skills as observed in the classroom was also included in the teacher rating form—the Work-Related Skills subscale of the Cooper-Farran Behavior Rating Scale (CFBR; Cooper & Farran, 1988). The 16 items on this scale ask about children's independent work, compliance with instructions, memory for instructions, and completion of games and activities. Items are rated from 1 to 7 using behavioral anchors distinctive to each item. Cronbach alphas were .95 at the beginning and end of pre-K.

## Academic Achievement Measures

Achievement was measured with five subscales from the Woodcock Johnson III achievement battery (Woodcock, McGrew, & Mather, 2001) widely used in early childhood education research. These included two math subtests: Applied Problems (numerical and spatial problems) and Quantitative Concepts (numbers, sequencing, shapes, and symbols). Language and literacy skills were assessed with Letter-Word Identification (identify and pronounce letters and read words), Picture Vocabulary (name objects in pictures and point to the picture that goes with a word), and Oral Comprehension (complete an orally presented passage by providing the appropriate missing word). Data analysis used the IRT-

scaled W-scores, but standard scores (mean of 100, standard deviation of 15) are more descriptive and showed fall pre-K baseline mean values for the pre-K sample of 98 on Applied Problems, 90 on Quantitative Concepts, 104 on Letter-Word Identification, 100 on Picture Vocabulary, and 97 on Oral Comprehension.

## Participants and Assessment Procedure

Parental consent was obtained for 608 children recruited from 58 pre-K classrooms in 32 schools/centers across four school systems and five community childcare centers in middle Tennessee. The consent rate was 60% (range 13% to 100% across classrooms). Consented children identified as English Language Learners were screened for English proficiency using the Pre-LAS (Duncan & DeAvila, 1985). Thirty-six children did not pass the Pre-LAS, 5 did not assent, and 32 moved before the study ended, leaving 535 children in the final analytic sample.

Participating schools/centers were in urban, suburban, and rural settings and provided a racially and economically diverse sample of children. Although information about race and economic status was not available for individual children, aggregate data for the schools/centers showed proportions of African American children that ranged from 0% to 87% ( $M = 16\%$ ), Hispanic children from 2% to 34% ( $M = 11\%$ ), and non-Hispanic White children from 13% to 95% ( $M = 71\%$ ). Economic diversity was indicated by a range of children qualifying for free or reduced price lunch programs from 16% to 100% ( $M = 55\%$ ). The children in the analytic sample ranged in age from 3.8 to 5.4 ( $M = 4.6$ ) at the beginning of pre-K and 52% were male.

**Procedure.** Children were assessed twice during the pre-K year—near the beginning (early September through October) and the end (mid-March to early May), referred to as Time 1 and Time 2, respectively. They were assessed again at the end of kindergarten (mid-March to early May; Time 3). Time 1 and 2 assessments were administered in three sessions of 20–30 min with nearly all sessions occurring within 10 or fewer weeks. Time 3 assessments were administered in two sessions spanning fewer than five days on average. Each child was assessed individually in a quiet area away from the classroom with a varying order for sessions but a fixed order for the measures within a session. In pre-K, the sessions included (a) Operation Span, Whisper, Peg Tapping, and WJ-III Applied Problems and Quantitative Concepts; (b) DCCS, HTKS, Digit Span, Copy Design, and WJ-III Picture Vocabulary; and (c) Spatial Conflict, Turtle-Rabbit, KRISP, and WJ-III Letter-Word Identification and Oral Comprehension. Based on the findings from pre-K, a reduced set of measures was administered in the two sessions at the end of kindergarten: (a) Peg Tapping, HTKS, Copy Design, and WJ-III Applied Problems, Quantitative Concepts, and Picture Vocabulary; and (b) DCCS, KRISP, Digit Span, and WJ-III Letter-Word Identification and Oral Comprehension.

Teacher ratings were made at approximately the same times as the child assessments near the beginning and end of the pre-K year. Kindergarten teachers then completed the same rating scales near the end of the kindergarten year.

**Missing data.** Of the 535 children who comprised the initial pre-K analytic sample, 47 could not be located for the Time 3 end of kindergarten assessments, leaving 488 children in the follow-up sample. The children missing Time 3 data were compared with



those providing data on the available demographic variables and the T1 and T2 CSR and achievement measures. *T* tests with Benjamini-Hochberg corrections for the large number of multiple comparisons showed no significant differences between children assessed and not assessed in kindergarten. Given no indications that the missing cases made the follow-up sample unrepresentative of the initial sample, analyses with pre-K data were conducted on the analytic sample of 535 children while those with kindergarten data were conducted on the sample of 488.

**Cross-validation sample and assessment procedure.** The cross-validation sample was drawn from a later cohort of children enrolled in pre-K in the four school systems that provided most of the original sample. These children were assessed three times during the pre-K year—near the beginning (Time 1), approximately 2 weeks later (retest) to assess the test–retest reliability of the measures, and near the end of the school year (Time 2). Parental consent was obtained for 593 children from 43 classrooms in 23 schools (overall consent rate of 69%). To accommodate limited resources for individual testing, only 10 consented children were randomly selected from classrooms with more than 10. This procedure produced a sample of 416 children, but 21 did not pass the Pre-LAS screen for English proficiency, four did not assent to the assessments, 18 moved prior to the reliability retest, and 4 were withdrawn due to assessor error. This left 369 children in the sample for the test–retest reliability data collected in the fall of the pre-K year. After that, 13 children moved before the end of pre-K, leaving 356 in the sample with data from both the beginning and end of the pre-K year.

The mean age of the children in both the test–retest and final samples was 4.4 years and 53% were male. As in the initial sample, the schools from which these children were drawn were economically and racially diverse: the proportion of students at each school qualifying for free or reduced price lunch ranged from 26% to 95% ( $M = 52\%$ ); the proportion who were African American ranged from 0% to 49% ( $M = 12\%$ ), the proportion Hispanic ranged from 1% to 38% ( $M = 9\%$ ), and the proportion non-Hispanic white ranged from 33% to 97% ( $M = 75\%$ ).

At Times 1 and 2, there were two assessment sessions, one for CSR and one for achievement. The order of these sessions varied, but the measures were administered in fixed order at each session. Only CSR measures were administered at Retest. In addition, at Time 1, Retest, and Time 2, teachers completed ratings on selected CSR measures (described later). The majority (74%) of these teachers had also participated in the initial phase of this study.

## Results

Analysis of the data described above was organized to compare the 12 candidate CSR measures with regard to their performance in the three areas described earlier that we judged to be especially pertinent to applications in pre-K settings where relevance to academic achievement is a major concern: (a) their predictive ability for academic achievement, (b) responsiveness to developmental change, and (c) concurrence with teacher ratings.

### Predictive Ability for Academic Achievement

The most important consideration for our purposes in assessing the CSR measures was their ability to predict academic achieve-

ment, measured here with the WJ-III Quantitative Concepts, Applied Problems, Oral Comprehension, Picture Vocabulary, and Letter-Word Identification subtests. The intercorrelations among these five subtests at Times 1, 2, and 3 were positive and relatively high, and principal components analyses showed strong one-factor solutions with loadings from .61 to .84. To represent overall academic achievement, therefore, we created a composite score for each time of measurement by combining the *W*-scores across the five subscales for each child with each subtest given equal weight.

**CSR predicting achievement.** The most direct answer to the question of the relative strength of the relation between each of the selected CSR measures and later achievement is obtained by comparing their correlations at each time of measurement with achievement measured at a later time. To address this question we first standardized the WJ composite achievement measure and each of the CSR measures separately for each time of testing so that the magnitude of the respective relations could be easily compared. We then constructed multilevel regression models in which each CSR measure in turn was used as the sole predictor of achievement at a later time. Multilevel analysis was necessary to respect the structure of the data and ensure that standard errors were properly estimated; it was conducted with SPSS 23 Mixed Models with children nested in classrooms, classrooms in schools (three levels), and both classrooms and schools treated as random effects. All the time intervals available in our data were examined: predicting from Time 1 (beginning of pre-K) to Time 2 (end of pre-K) and Time 3 (end of kindergarten), and predicting from Time 2 to Time 3.

Table 1 reports the standardized regression coefficients estimated in each of these analyses. Because all the variables were standardized and there was only one predictor in each analysis, these coefficients can be read as zero-order product-moment correlation coefficients. All these correlations were statistically significant with the largest found for Backwards Digit Span, Copy Design, DCCS, HTKS, KRISP Accuracy, and Peg Tapping. From the beginning of pre-K to the end of pre-K (Time 2) and then to the end of kindergarten (Time 3), the correlations for those CSR measures ranged from .37 to .56. From the end of pre-K (Time 2) to the end of kindergarten, they ranged from .38 to .55.

**CSR predicting achievement gain.** The analyses reported above show that children with better initial skills on the CSR measures show higher achievement levels at a later time, but those children also have higher achievement to begin with—the concurrent CSR–achievement correlations at Time 1 and Time 2 for the best CSR measures in Table 1 ranged from .42 to .59. The next set of analyses therefore addressed the further question of the relative ability of the CSR measures to predict the achievement gains made over a subsequent period. Our interest is in achievement gains associated with the experiences children have over a school year, not the portion predictable from their initial achievement levels prior to those experiences. For these analyses, we used the same 3-level regression models described above, but with the Time 1 WJ composite variable included as a covariate in each analysis along with the respective CSR measure. The CSR measures in these analyses, therefore, were predicting residual gain in achievement; that is, later achievement with initial achievement held constant (Cronbach & Furby, 1970).

Table 1  
Standardized Regression Coefficients Between Each of the Cognitive Self-Regulation (CSR) Measures and Later Academic Achievement for the Initial and Cross-Validation Samples

CSR measure	Initial sample (n = 535)			Cross-validation sample (n = 356)
	Time 1 CSR & Time 2 Achievement	Time 1 CSR & Time 3 Achievement	Time 2 CSR & Time 3 Achievement	Time 1 CSR & Time 2 Achievement
Backwards Digit Span	.42	.37	.47	.46
Copy Design	.41	.40	.38	.40
DCCS	.45	.44	.42	.50
HTKS	.50	.49	.55	.52
KRISP Accuracy	.48	.50	.43	.46
KRISP Reaction Time	.25	.23	.21	— <sup>a</sup>
Operation Span	.26	.27	.21	—
Peg Tapping	.56	.51	.52	.58
Spatial Conflict	.29	.27	.18	—
Turtle-Rabbit Accuracy	.22	.23	.18	—
Turtle-Rabbit Reaction Time	.32	.31	.26	—
Whisper Task	.37	.36	.25	—
CSR factor score				.79

Note. N = 488 at Time 3. All correlations are statistically significant at  $p < .01$  in multilevel analysis. DCCS = Dimensional Change Card Sort; HTKS = Head Toes Knees Shoulders; KRISP = Kansas Reflection-Impulsivity Scale for Preschoolers. Academic achievement is the composite measure combining five Woodcock-Johnson subscales. Time 1 = beginning of pre-K; Time 2 = end of pre-K; Time 3 = end of kindergarten. The CSR factor score is based on the six individual CSR measures shown for the cross-validation sample.

<sup>a</sup> Measure not included in cross-validation.

The first three columns of Table 2 show standardized regression coefficients from these analyses. It is not surprising that they are relatively small given the strong relation between initial and later achievement. Nonetheless, many of the CSR measures had statistically significant predictive relations with achievement gain from the beginning to the end of pre-K and to the end of kindergarten, as well as from the end of pre-K to the end of kindergarten. The measures with significant positive predictive relations for all three intervals, at least at  $p < .10$ , were Backwards Digit Span, Copy Design, HTKS, KRISP Accuracy, and Peg Tapping.

Table 2  
Standardized Regression Coefficients for the Relation Between Each Cognitive Self-Regulation (CSR) Measure and Residual Gain on the Academic Achievement Composite for the Initial and Cross-Validation Samples

CSR measure	Initial sample (n = 535)					Cross-validation sample (n = 356)	
	Time 1 CSR & T1-T2 Ach Gain	Time 1 CSR & T1-T3 Ach Gain	Time 2 CSR & T2-T3 Ach Gain	T1-T2 CSR Gain & T1-T2 Ach Gain	T1-T2 CSR Gain & T1-T3 Ach Gain	Time 1 CSR & T1-T2 Ach Gain	T1-T2 CSR Gain & T1-T2 Ach Gain
Backwards Digit Span	.06*	.05 <sup>†</sup>	.08*	.12*	.14*	.05	.06 <sup>†</sup>
Copy Design	.12*	.12*	.05*	.07*	.06*	.05	.10*
DCCS	.07*	.10*	.04	.10*	.06*	.11*	.10*
HTKS	.10*	.08*	.13*	.09*	.14*	.06 <sup>†</sup>	.08*
KRISP Accuracy	.09*	.17*	.10*	.09*	.08*	.09*	.09*
KRISP Reaction Time	.09*	.09*	.02	.05*	.05 <sup>†</sup>	— <sup>a</sup>	—
Operation Span	.07*	.09*	.01	.05*	.02	—	—
Peg Tapping	.09*	.09*	.05 <sup>†</sup>	.11*	.07*	.10*	.15*
Spatial Conflict	.06*	.06*	.03	.05*	.05 <sup>†</sup>	—	—
Turtle-Rabbit Accuracy	.03	.05 <sup>†</sup>	−.02	.08*	.03	—	—
Turtle-Rabbit Reaction Time	.03	.05 <sup>†</sup>	.07*	.01	.04	—	—
Whisper Task	.06*	.07*	−.05*	.09*	−.01	—	—
CSR factor score						.18*	.16*

Note. N = 488 at T3. Ach = Achievement; DCCS = Dimensional Change Card Sort; HTKS = Head Toes Knees Shoulders; KRISP = Kansas Reflection-Impulsivity Scale for Preschoolers; RT = Reaction Time. Academic achievement is the composite measure combining five Woodcock-Johnson subscales. Time 1 = beginning of pre-K; Time 2 = end of pre-K; Time 3 = end of kindergarten. The CSR factor score is based on the 6 individual CSR measures shown for the cross-validation sample.

<sup>a</sup> Measure not included in cross-validation.

<sup>†</sup>  $p < .10$ . \*  $p < .05$ .



**CSR gain predicting achievement gain.** The last set of analyses addressing CSR-achievement relations compared the CSR measures with regard to the extent to which the CSR skill gains they showed during the pre-K year, and between the end of pre-K and end of kindergarten, were correlated with achievement gains made over those same periods. For this gain-with-gain analysis, we first used the same three-level regression models described earlier to estimate residual gain for each CSR measure over the respective periods by predicting later CSR scores from the initial (Time 1) values on the same CSR measure. The residuals from those analyses, representing the changes in the CSR measure that cannot be predicted from their initial status, are residual gain scores for the CSR measures. Those residual gain scores were then used as independent variables in a second series of multilevel regression analyses in which each CSR residual gain score was used to predict later achievement with initial achievement controlled, the analysis model used above to examine residual gain on achievement.

The fourth and fifth columns of Table 2 report the standardized regression coefficients from these analyses. As in the previous analysis, these coefficients are relatively small because the much larger relations between prepost CSR and prepost achievement have been adjusted out of the results. The relations of CSR residual gain during pre-K with residual achievement gain during that same year, and with residual achievement gain between the beginning of pre-K and the end of kindergarten, are nonetheless statistically significant for many of the CSR measures. The better performing CSR measures across these various intervals, as indicated by the pattern of statistical significance, were Backwards Digit Span, Copy Design, DCCS, HTKS, KRISP Accuracy, and Peg Tapping.

Responsiveness to Developmental Change

The last set of analyses reported above demonstrated that residual gain on some of the CSR measures was significantly related to residual gain on the achievement measures. However, those analyses do not directly address the question of how much change there is on each CSR measure during the pre-K year. As noted

earlier, the most educationally relevant CSR measures are those capable of showing the most growth during the pre-K year. To examine the responsiveness of the measures to developmental change, children’s scores on each CSR measures at the beginning of pre-K were compared to their scores at the end of the year. These analyses were conducted with four-level regression models in which a dummy code for time predicted each CSR score with Time 1 and Time 2 scores nested within children and children nested within classrooms and schools. The CSR scores were not standardized for this analysis, allowing estimation of the mean scores at Time 1 (time = 0) and Time 2 (time = 1) in the original metric. Table 3 shows the means and the standard deviations for each CSR measure. The difference between children’s performance at Time 1 and Time 2, indexed by the regression coefficient on the time dummy code, was statistically significant for all the CSR measures except Turtle-Rabbit Accuracy. Pre-post standardized mean difference effect sizes are also shown in Table 3, computed as the Time 2 mean minus the Time 1 mean divided by the pooled standard deviation. These effect sizes for all the measures other than Turtle-Rabbit accuracy were positive and ranged from .31 to .69, with the greatest gains for Copy Design, DCCS, HTKS, KRISP Accuracy, and Peg Tapping (effect sizes greater than 0.50).

Table 3 also shows the zero-order product-moment correlations between children’s scores at the beginning and end of pre-K. These were all statistically significant and ranged from .12 to .66. The largest of them showed reasonable consistency in children’s relative ranking over the school year. Nevertheless, they were not so large as to indicate that only stable individual differences are reflected in these CSR measures with no room for influence from differential experiences in and out of the classroom during this period.

Concurrence With Teacher Ratings

To investigate the relation between the CSR measures and teacher’s ratings of CSR-related behavior in the classroom, we examined the correlations between each CSR measure and each of the five teacher rating scales (CFBR Work Related Skills,

Table 3  
*Change in Scores on the Cognitive Self-Regulation (CSR) Measures from the Beginning (Time 1) to End of Pre-K (Time 2)*

CSR measure	Time 1: <i>M (SD)</i>	Time 2: <i>M (SD)</i>	T1 — T2 effect size	T1 — T2 correlation
Backwards Digit Span	1.31 (1.20)	2.05 (2.13)	.43	.46
Copy Design	1.40 (1.43)	2.27 (1.70)	.55	.59
DCCS	1.47 (.57)	1.75 (.52)	.51	.38
HTKS	8.91 (11.89)	15.51 (14.11)	.51	.61
KRISP Accuracy	28.94 (4.09)	31.44 (3.13)	.69	.56
KRISP Reaction Time	.15 (.34)	.30 (.26)	.50	.12
Operation Span	8.57 (3.87)	9.67 (3.18)	.31	.38
Peg Tapping	6.99 (6.01)	10.21 (5.48)	.56	.62
Spatial Conflict	20.86 (6.82)	22.82 (6.06)	.30	.31
Turtle-Rabbit Accuracy	54.18 (9.96)	54.22 (6.62)	.00	.20
Turtle-Rabbit Reaction Time	5.84 (8.33)	10.69 (15.43)	.39	.54
Whisper Task	30.04 (8.13)	32.82 (6.04)	.39	.35

*Note.* *N* = 535. The pre-post difference is statistically significant at *p* < .001 for all measures except Turtle-Rabbit Accuracy. Effect size is Cohen’s *d* for the difference between the means at Time 1 and Time 2. DCCS = Dimensional Change Card Sort; HTKS = Head Toes Knees Shoulders; KRISP = Kansas Reflection-Impulsivity Scale for Preschoolers.

TABC Distractibility, TABC Persistence, CBQ Attention Shifting, and CBQ Impulsivity) and a composite scale created by summing z-scores computed for each teacher rating scale. These correlations were estimated as standardized regression coefficients in 3-level regression models in which the respective CSR measure at either Time 1 or Time 2 was the sole predictor of a teacher rating obtained at the corresponding time. The correlations of each CSR measure with the composite scale and with each individual teacher rating scale are reported in Table 4 for the beginning and end of pre-K.

As Table 4 shows, all these correlations were statistically significant except for a few involving CBQ Impulsivity. The largest correlations with the Teacher Rating Composite appeared for Peg Tapping, HTKS, and KRISP Accuracy (.34 to .42). Close behind were Copy Design, DCCS, and Turtle-Rabbit Accuracy with correlations of at least .25. The correlations were substantially similar for ratings at the beginning and end of pre-K. The correlations with individual teacher rating scales showed similar patterns, though lower for the CBQ scales.

Summary of Findings on the Selected Criteria

Table 5 summarizes the comparative findings reported above for the performance of the candidate CSR measures by identifying the top performers in each analysis based on the magnitude of the parameter estimates and/or statistical significance. The measures are listed with the better performing ones first rather than in alphabetical order as in the previous tables. Four CSR measures were among the top performers in every analysis: Copy Design, HTKS, KRISP Accuracy, and Peg Tapping. DCCS was very close behind, appearing in the top performing group in all but one analysis. Consideration must also be given

to Backwards Digit Span, which showed good performance for predicting achievement, though it was not among the top performers in the other analyses. The most notable feature of this summary is the consistency of the CSR measures that performed well—those that were strong in one analysis were strong in all or nearly all of them, and those weak in any one analysis were weak in all or nearly all.

Performance of the Top CSR Measures in Combination

As the summary in Table 5 indicates, there were six CSR measures that performed best in the comparative analyses. With those results in hand, we then undertook an exploration of the relations of those six measures to achievement when taken altogether to determine which showed the strongest independent relations relative to the others and to assess the potential value of a composite of multiple measures. For that purpose, another series of three-level regression analyses was conducted with all six of these measures used together as predictors. To examine their collective performance, multiple correlations were estimated for their relations to the different dependent variables of interest. This was done by first fitting the models with the six CSR measures omitted to obtain an estimate of the total unconditional variance (residual variance when the achievement pretest was a necessary covariate) across all levels on the respective dependent variables. We then ran the same models with all six measures included as predictors and obtained the total conditional variance from those analyses. The difference between the total unconditional variance without the six CSR measures and the total conditional variance with them in the model represents the amount of the total between-student

Table 4  
Concurrent Correlations Between Child Cognitive Self-Regulation (CSR) Measures and Teacher Rating Scales at the Beginning (Time 1) and End of Pre-K (Time 2) for Initial and Cross Validation Samples

CSR measure	Initial sample (n = 535)										Cross-validation sample (n = 356)			
	CFBR—work related skills		TABC— distractibility		TABC— persistence		CBQ—attention shifting		CBQ— impulsivity		Teacher rating composite		Teacher rating total score	
	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2	Time 1	Time 2
Backwards Digit Span	.24	.20	.22	.15	.22	.17	.14	.14	.04	.05	.21	.17	.30	.27
Copy Design	.34	.32	.33	.28	.31	.32	.18	.20	.11	.08	.31	.29	.42	.47
DCCS	.27	.25	.29	.24	.27	.25	.21	.17	.13	.12	.28	.25	.36	.30
HTKS	.36	.39	.35	.39	.28	.40	.28	.29	.10	.18	.34	.39	.39	.41
KRISP Accuracy	.38	.39	.35	.35	.32	.38	.28	.28	.12	.24	.36	.40	.41	.38
KRISP RT	.22	.20	.19	.13	.21	.16	.16	.13	.01	.03	.19	.16	— <sup>a</sup>	—
Operation Span	.23	.19	.23	.14	.15	.16	.14	.13	.06	.08	.20 <sup>†</sup>	.17	—	—
Peg Tapping	.43	.39	.42	.36	.36	.36	.34	.28	.16	.19	.42	.38	.45	.41
Spatial Conflict	.18	.15	.24	.17	.19	.20	.17	.13	.16	.12	.23	.19	—	—
Turtle-Rabbit Accuracy	.24	.23	.27	.26	.20	.25	.23	.23	.13	.18	.26	.28	—	—
Turtle-Rabbit RT	.21	.17	.19	.13	.16	.13	.12	.08	.03	.06	.17	.14	—	—
Whisper Task	.27	.17	.27	.19	.19	.20	.22	.10	.03	.14	.24	.20	—	—

Note. Correlations greater than .09 are statistically significant at  $p < .05$  in multilevel analysis. DCCS = Dimensional Change Card Sort; HTKS = Head Toes Knees Shoulders; KRISP = Kansas Reflection-Impulsivity Scale for Preschoolers; RT = reaction time.  
<sup>a</sup> Measure not included in cross-validation.



Table 5

*Summary of the Performance of the Candidate Cognitive Self-Regulation (CSR) Measures on the Attributes Examined*

CSR measure	Predicting achievement				
	T1 & T2 CSR & later achievement <sup>a</sup>	T1 & T2 CSR & achievement gains <sup>b</sup>	PreK CSR gains & achievement gains <sup>c</sup>	Developmental change <sup>d</sup>	Concurrence with teacher ratings <sup>e</sup>
Copy Design	X	X	X	X	X
HTKS	X	X	X	X	X
KRISP Accuracy	X	X	X	X	X
Peg Tapping	X	X	X	X	X
DCCS	X		X	X	X
Backwards Digit Span	X	X	X		
Turtle-Rabbit Accuracy					X
KRISP Reaction Time					
Operation Span					
Turtle-Rabbit Reaction Time					
Spatial Conflict					
Whisper Task					

*Note.* The better performing CSR measures on each criterion are indicated by X. DCCS = Dimensional Change Card Sort; HTKS = Head Toes Knees Shoulders; KRISP = Kansas Reflection-Impulsivity Scale for Preschoolers. Time 1 (T1) = beginning of pre-K; Time 2 (T2) = end of pre-K; Time 3 (T3) = end of kindergarten.

<sup>a</sup> Correlations for T1 predicting to T2 and T3 achievement, and T2 predicting to T3 achievement are  $\geq .35$  and significant at  $p \leq .05$ . <sup>b</sup> Correlations for T1 predicting T1-T2 and T1-T3 achievement gain, and T2 predicting T2-T3 achievement gain are significant at  $p \leq .10$  or better. <sup>c</sup> Correlations for T1-T2 gain predicting T1-T2 and T1-T3 achievement gain are significant at  $p \leq .05$ . <sup>d</sup> Effect size for change from T1 to T2 is  $\geq .50$ . <sup>e</sup> T1 and T2 correlations with the Teacher Rating Composite are  $\geq .25$  and significant at  $p \leq .05$ .

variance accounted for by the CSR measures, essentially an R-squared value when represented as a proportion. The square root of that estimate was taken as the multiple correlation of interest. In addition, the standardized regression coefficient for each CSR measure indicated the independent contribution that measure made to predicting the respective dependent variable.

The results of these analyses are summarized in the upper portion of Table 6. The first panel reports the collective relation of the six CSR measures to composite achievement measured later. The multiple correlations, ranging from .68 to .72, can be compared with the analogous correlations for the individual measures shown in Table 1, all of which are smaller. The standardized regression coefficients indicate that the strongest independent contributions were made by HTKS, KRISP Accuracy, Peg Tapping, and Backwards Digit Span.

The second panel of Table 6 provides the results for the six CSR measures collectively predicting achievement gain over various periods. The multiple correlations, ranging from .23 to .28, can be compared with the standardized regression coefficients in the first four columns of Table 2, all of which are notably smaller. The regression coefficients in Table 6 indicate that KRISP Accuracy and HTKS have the strongest independent relations to achievement gain, followed by Copy Design and Backwards Digit Span.

The third panel in Table 6 reports the results for the most important predictive relations with achievement—those between residual gain on the CSR measures and residual achievement gain. The multiple correlations, which can be compared with the smaller standardized regression coefficients for the individual measures in the last three columns of Table 2, ranged from .32 to .39. The individual measures making the strongest independent contributions were Backwards Digit Span and Peg Tapping.

The last panel of Table 6 shows the multiple correlations that index the concurrent relations of the set of six CSR measures with the composite teacher ratings at the beginning (T1) and end (T2) of pre-K. Those multiple correlations (.49 and .50, respectively) can be compared with the analogous correlations for each individual CSR measure reported in the first two columns of Table 4, all of which are smaller. The standardized regression coefficients in Table 6, in turn, indicate that Peg Tapping, KRISP Accuracy, and HTKS made the strongest independent contributions to those relations.

The results in Table 6 show, unsurprisingly, that a combination of the six top performing individual CSR measures has greater predictive relations with achievement and more concurrence with teacher ratings than any single measure. Moreover, in most instances the improvement in the magnitude of the respective relations is great enough to indicate that a composite of measures holds more promise as a general measure of CSR for pre-K children than any one of them used alone. Among the six measures, the strongest independent contributions were made by Peg Tapping, KRISP Accuracy, and HTKS, which would thus be the leading candidates for the most efficient composite measure. In addition, Backwards Digit Span had an especially strong influence in the relation between CSR gain and achievement gain and thus would deserve some consideration as well.

### Cross-Validation

As described above, six of the candidate CSR child assessments performed well in our comparative analyses. However, the large number of analyses conducted to identify those six allow ample opportunity for chance factors in the particular sample of children and the data they provided to influence the results. In the follow-up cross-validation study, therefore, we administered those six mea-

Table 6  
Multiple Correlation and Regression Coefficients for the Top Six Cognitive Self-Regulation (CSR) Measures Together Predicting Achievement and Concurring With Teacher Ratings in the Initial Sample (Top Panel) and Cross-Validation Sample (Bottom Panel)

IVs and DV	Multiple correlation	Standardized regression coefficients for CSR measures					Backwards Digit Span
		Copy Design	HTKS	KRISP accuracy	Peg tapping	DCCS	
Initial sample ( <i>N</i> = 535)							
IVs: T1 CSR measures DV: T2 achievement	.72*	.10*	.17*	.19*	.24*	.17*	.19*
IVs: T1 CSR measures DV: T3 achievement	.68*	.10*	.11*	.25*	.19*	.19*	.17*
IVs: T2 CSR measures DV: T3 achievement	.71*	.07*	.25*	.17*	.19*	.12*	.23*
IVs: T1 CSR measures DV: T1–T2 achievement gain	.29*	.08*	.05*	.05*	.04	.04 <sup>†</sup>	.05 <sup>†</sup>
IVs: T1 CSR measures DV: T1–T3 achievement gain	.30*	.06*	.02	.14*	.03	.07*	.04
IVs: T2 CSR measures DV: T2–T3 achievement gain	.25*	.02	.10*	.08*	.01	.01	.06*
IVs: T1–T2 CSR gain DV: T1–T2 achievement gain	.38*	.04*	.06*	.06*	.08*	.07*	.10*
IVs: T1–T2 CSR gain DV: T1–T3 achievement gain	.32*	.04	.11*	.05*	.04	.04	.12*
IVs: T1 CSR measures DV: T1 teacher ratings	.56*	.16*	.09*	.15*	.25*	.08 <sup>†</sup>	.02
IVs: T2 CSR measures DV: T2 Teacher ratings	.57*	.11*	.24*	.22*	.19*	.02	−.04
Cross-Validation Sample ( <i>N</i> = 356)							
IVs: T1 CSR measures DV: T2 achievement	.73*	.02	.13*	.22*	.25*	.22*	.19*
IVs: T1 CSR measures DV: T1–T2 achievement gain	.26*	−.01	.00	.08*	.06 <sup>†</sup>	.10*	.04
IVs: T1–T2 CSR Gain DV: T1–T2 achievement gain	.37*	.08*	.05 <sup>†</sup>	.06*	.11*	.07*	.04
IVs: T1 CSR measures DV: T1 teacher ratings	.60*	.19*	.10 <sup>†</sup>	.19*	.21*	.08	.04
IVs: T2 CSR measures DV: T2 teacher ratings	.62*	.30*	.16*	.18*	.11 <sup>†</sup>	.10*	.03

Note. IV = independent variable; DV = dependant variable; DCCS = Dimensional Change Card Sort; HTKS = Head Toes Knees Shoulders; KRISP = Kansas Reflection-Impulsivity Scale for Preschoolers. Academic achievement is the composite measure combining five Woodcock-Johnson subscales. Time 1 (T1) = beginning of pre-K; Time 2 (T2) = end of pre-K; Time 3 (T3) = end of kindergarten.  
<sup>†</sup>  $p < .10$ . \*  $p < .05$ .

tures<sup>1</sup> to a new sample to check the stability of the key features that favored them in the initial analyses. We also used this new sample to estimate the test–retest reliability of the selected measures.

The six selected CSR measures and the WJ-III achievement measures used in the initial phase were administered in two sessions at the beginning (Time 1) and end (Time 2) of the pre-K year. The order of these sessions was varied, but the measures were administered in a fixed order at each session. The CSR measures were administered a second time approximately two and a half weeks after the assessment sessions at the beginning of the year to allow estimation of test–retest reliability. In addition, at Time 1, Retest, and Time 2, teachers completed ratings on a subset of 20 teacher ratings items from the initial phase: 10 items from CFBR Work-Related Skills, 3 from CBQ (2 Impulsivity, 1 Attention Shifting), and 7 from TABC (3 Persistence, 4 Distractibility). The 20 selected items were those that had the largest loadings with the common factor identified by a principal components analysis of the original 57 items. Factor loadings greater than .70 for data collected at both the beginning and end of pre-K indicated that

these 20 items efficiently represented the principal factor underlying the original 57 items and they were therefore used in the cross-validation to reduce the response burden on the teachers. These items were all rated on 7-point scales and showed a high level of internal consistency (Cronbach’s alpha of .98 at both Time 1 and 2). A total score was computed as the mean of the 20 items.

**Test–retest reliability.** The mean interval between the CSR assessments for the 369 children in the test–retest sample was 16.7 days ( $SD = 5.0$ ). Test–retest reliability was estimated using multilevel regression to account for the effect of the nesting of children within classrooms and classrooms within schools. For each mea-

<sup>1</sup> Scores for the Backward Digit Span measure in the cross-validation reflect the longest span correctly recalled (range = 1–8) based on administration procedures from the Wechsler Intelligence Scale for Children, 4th edition (Wechsler, 2003). For the KRISP, we added more advanced items from version B to provide a better ceiling (maximum score of 48). Scoring was altered for Copy Design; every attempt was scored of the two allowed for each item, making the scores range from 0 to 16. The other cognitive self-regulation measures were the same as before.



sure, the initial score was used to predict the retest score with the standardized regression coefficients then representing test-retest correlations. In descending order, those reliability coefficients and their standard errors were Peg Tapping, .80 (.03); HTKS, .78 (.03); Backwards Digit Span, .73 (.04); Copy Design, .72 (.04); KRISP Accuracy, .64 (.04); and DCCS, .47 (.05). The KRISP reliability coefficient is modest and that for DCCS is marginal, but the others are in a generally acceptable range. Test-retest reliability was also estimated for a composite of all six measures, yielding a reliability coefficient of .89 (.02).

**Predictive relations with achievement.** To assess the ability of the CSR measures to predict academic achievement in the cross-validation sample, we first examined the correlations between the CSR measures at Time 1 and the composite achievement score at Time 2 (column 4, Table 1). As with the initial sample, these were estimated with standardized regression coefficients in multilevel models. These coefficients were statistically significant and very similar to those found in the initial sample (column 1, Table 1). The correlation with achievement for a composite score that combined all six CSR measures is also shown in Table 1 and demonstrates that the combined set of items performs notably better than any one item.

The ability of each CSR measure to predict the gain children in the cross-validation sample made in achievement over the pre-K year was also assessed with standardized regression coefficients from multilevel models in which Time 1 achievement was controlled. These coefficients were statistically significant for DCCS, HTKS, KRISP Accuracy, and Peg Tapping (column 6, Table 2), but showed some modest inconsistencies with the initial sample results (column 1, Table 2) for Copy Design (.05 vs. .12), DCCS (.11 vs. .07), and HTKS (.06 vs. .11). The coefficients for the more revealing relations between gains on the CSR measures and gains in achievement over the pre-K year (Time 1 to Time 2) were statistically significant for all the measures (column 7 in Table 2). The strongest relations were for Peg Tapping, DCCS, Copy Design, and KRISP Accuracy but here also there were some modest inconsistencies with the estimates from the original sample (column 5, Table 2) for some measures, specifically Backwards Digit Span (.06 vs. .12), Copy Design (.10 vs. .07), and Peg Tapping (.15 vs. .11). The predictive coefficients for the composite score that combined all six CSR measures are also shown in Table 2 and here also the combined set of items performs better than any one item.

The final series of analyses for the predictive relations with achievement in the cross-validation sample investigated the independent contribution of each of the six CSR measures relative to the others when they were used simultaneously as independent variables in multilevel regressions. These analyses were conducted using the same models and procedures described earlier for the analogous analyses with the initial data. The results are reported in Table 6 (bottom panel) along with those for the initial sample (top panel). Across all outcomes, Peg Tapping, DCCS, and KRISP Accuracy showed the largest independent relations to achievement and these were the only three CSR measures for which the coefficients were statistically significant with every outcome. However, Copy Design showed a significant independent gain-with-gain relation and HTKS and Backwards Digit Span showed significant independent contributions to predicting Time 2 Achievement. Comparing these results with the analogous ones for the initial sample, the coefficients most similar on statistical sig-

nificance and magnitude across the achievement outcomes were for KRISP Accuracy, though Peg Tapping, DCCS, and HTKS also showed relatively good replication.

**Concurrent relations with teacher ratings.** The total score of the 20 teacher rating items was used as the dependent variable in multilevel regression models with each CSR measure in turn as the sole independent variable, as in the analogous analysis with the initial sample. The standardized regression coefficients that represent the correlations between each CSR measure and the teacher rating total score are reported in the columns on the far right in Table 4. They ranged from .27 to .47 and all were statistically significant. The largest correlations were found for Copy Design, Peg Tapping, HTKS, and KRISP Accuracy. Compared with the analogous values from the initial sample (also shown in Table 4), all but two of these correlations are larger and those two are close to the prior values. A broader view is provided by the correlations between the total teacher rating scores at Times 1 and 2 and the composite score that combined all six CSR measures. These were .54 and .60, respectively (not shown in Table 4), again showing stronger relations than any of the individual measures in that composite.

## Useful Information

As shown in analyses with both the initial and cross-validation samples, none of the individual top performing CSR measures did nearly as well in our tests as a composite of all six of them. Moreover, each of the six measures made an independent contribution to the predictive strength of the composite, so no more efficient subset of fewer than all six measures would perform quite as well. Our procedure for administering those measures with the cross-validation sample demonstrated that it was feasible to include them in a single assessment session of 35–45 min. Further information about these six measures and how they are administered can be obtained from the corresponding author. It might be tempting to shorten the battery by omitting Copy Design and Backwards Digit Span, but analyses not reported here across both samples showed that this produced a notable decrement in the performance of the composite for predicting achievement. The six measures are scored on quite different scales, however, complicating the integration of them into a single composite measure. For research purposes, computing standardized z-scores for each, then summing them provides a straightforward way to create such a composite measure. Such standardization, however, makes the scoring dependent on the means and standard deviation of the particular sample on which the data were collected. Those values may not be well estimated in small samples and, in any event, such sample dependence undermines comparability across samples and studies. For more general use, each measure can be rescaled into a 0- to 5-point format with all six then summed to create a simple additive total score that works well. Appendix B describes the rationale, procedure, and results of this rescaling.

Use of any of the CSR measures identified in this study as outcome variables in research on the effects of interventions with pre-K students will likely involve cluster-randomized trials with students nested within classrooms and schools. The intraclass correlations (ICCs; also known as intracluster correlations) that characterize the proportions of total variance that are between schools and between classrooms within schools are critical for



estimating statistical power during the planning stage and influence the standard errors in multilevel analysis. The multilevel structure of the samples used in the present study allows ICCs to be estimated for classroom and school clusters. These estimates are reported in Table 7 for the initial sample, the larger of the two available, at the beginning of pre-K. They were estimated in three-level unconditional models with each of the CSR measures in turn as the dependent variable. The respective ICC values were computed as the proportion of the total variance associated with each of the levels in the multilevel structure. The between-school and between-classroom-within-school ICCs were relatively modest for this pre-K sample with the between-classroom value virtually zero for several measures.

Discussion

The objective of this study was to identify direct assessment measures of CSR for pre-K children that are closely linked to their academic achievement, that is, learning-related cognitive self-regulation (LRCSR), and that perform well on other criteria that make them educationally relevant for research and practical applications. In pursuing this objective, we evaluated existing measures in a comparative fashion, choosing candidate measures with attention to the aspect of CSR most salient in the tasks the measure presented to children, prior evidence about their association with academic achievement, and the ease with which they could be administered in classroom settings. The most important consideration for identifying the best performing of the selected measures was their ability to predict children’s subsequent academic achievement and achievement gains. Because an important use of such measures is as outcomes for research on interventions aimed at improving LRCSR, we also considered the extent to which the measures showed change over the pre-K year, thus demonstrating their ability to respond to increases in children’s LRCSR skills. Finally, we attended to the concurrent relations between the candidate measures and ratings by pre-K teachers of children’s LRCSR-related behavior in the classroom as a further indication of their educational relevance.

Table 7  
*Intraclass Correlation Coefficients Associated With Students Nested Within Classrooms and Classrooms Nested Within Schools in the Initial Sample*

CSR measure	Between schools	Between classrooms within schools	Between students within classrooms
Backwards Digit Span	.020	.000	.980
Copy Design	.049	.017	.934
DCCS	.028	.038	.934
HTKS	.006	.036	.958
KRISP			
Accuracy	.013	.000	.987
Peg Tapping	.035	.000	.965
CSR Composite Score	.027	.020	.953

*Note.* *N* = 535. CSR = cognitive self-regulation; DCCS = Dimensional Change Card Sort; HTKS = Head Toes Knees Shoulders; KRISP = Kansas Reflection-Impulsivity Scale for Preschoolers. The CSR Composite Score combines the six individual CSR measures shown.

Of the 12 candidate measures evaluated, analyses with the initial pre-K sample identified six that performed especially well against these criteria. Cross-validation with a new sample confirmed the stability of those findings. The best performing measures overall were Copy Design, HTKS, KRISP Accuracy, Peg Tapping, DCCS, and Backwards Digit Span. The single best performing measure across all our analyses was Peg Tapping, with the functionally similar HTKS close behind and KRISP Accuracy in third place.

The findings reported here complement and extend the body of research on the psychometric characteristics of CSR measures for pre-K age children. While same day test-retest reliability for DCCS has been documented (Beck, Schaefer, Pang, & Carlson, 2011) and normed performance standards have been established (Weintraub et al., 2013), the present study adds to the validation of this measure for use in preschool settings to assess learning-related CSR by demonstrating its relation with concurrent teacher ratings of CSR in situ and with later academic achievement. Also, with regard to research with the HTKS task, the present work adds to the somewhat mixed results from attempts to demonstrate associations between teacher ratings of classroom behavioral regulation and academic achievement (Graziano et al., 2015; McClelland et al., 2007; Ponitz et al., 2009) by demonstrating test-retest reliability in both CSR tasks and teacher ratings. Similarly, this study contributes test-retest and internal consistency reliability estimates and supportive validity data for preschool applications for all six of the measures that performed best in our comparative evaluation.

Although sound psychometric characteristics are fundamental for any CSR measure that will be used for practical or research purposes, the unique contribution of this study is the head-to-head comparison of the selected measures on a range of probing performance indicators related to their educational relevance for pre-K children in classroom settings. The results provide a firm empirical basis for the use of any of the top performing measures for either of the applications that motivated this study. The better performing measures showed close relations to achievement and achievement gains, sensitivity to developmental change, and reasonable congruence with the CSR-related behavior teachers observed in the classroom. These characteristics make them especially suitable as screening measures to identify children whose CSR skills may be low enough to impair their learning in pre-K contexts and to monitor improvement in those skills during the pre-K year. Further, those characteristics make the top performing measures suitable choices as outcomes for intervention research aimed at improving those CSR skills that have sufficiently close relations to learning that such improvement may, in turn, boost academic achievement. It is especially fortuitous in this regard that the best performing individual CSR measures are among the easiest to administer and score. Most notably, peg tapping and HTKS performed very well by the criteria we applied and both can be administered quickly and easily without special equipment or materials and without extensive training. Each could thus be used on a stand-alone basis with the results reported here providing assurance of their educational relevance for pre-K students.

Similar to Willoughby et al. (2016), however, we found that the best performing CSR measures work better as a composite. It is hardly surprising that a combination of related measures performs better than any individual measure by itself. What was somewhat unexpected was that each of the six made significant independent contributions to at least some of the relations examined with them



in combination. The best composite measure based on these results, therefore, would include all six individual measures. However, there were differences in the independent contribution each measure made to the performance of that composite, with Peg Tapping, KRISP Accuracy, and HTKS demonstrating the strongest independent relations, and gain on Backwards Digit Span showing an especially strong independent relation to achievement gain. A more efficient composite measure incorporating the three top performers in this analysis, possibly with Backward Digit Span included as a fourth, therefore, would also perform better than any single measure while not requiring data collection on all six.

It is notable that the six measures contributing to the full composite measure represented a mix of CSR skills—two primarily emphasizing sustained attention (Copy Design, KRISP), two emphasizing inhibitory control (Peg Tapping, HTKS), one emphasizing attention shifting (DCCS), and one emphasizing working memory (backward digit span). Although the skills in this mix were interrelated and all were found to be relevant to learning, no one skill dominated so strongly that the others were irrelevant. This grouping of tasks as indicators of CSR aligns with prior work supporting a multidimensional view of executive function (Miyake et al., 2000).

Recognition of the advantages of an ensemble of measures to fully represent various forms of CSR is not unique to the present study. The work on the NIH Toolbox of brief measures for executive function (Weintraub et al., 2013; Zelazo & Bauer, 2013) and the program of research by Willoughby and colleagues (Willoughby, Blair, Wirth, & Greenberg, 2012; Willoughby, Wirth, & Blair, 2011) also takes this approach. Moreover, Willoughby et al. (2016) identified a cluster of measures of executive function that are associated with academic achievement. Similarly, the Chicago School Readiness Project has developed a brief direct assessment battery of CSR measures appropriate for field-based settings (Smith-Donald, Raver, Hayes, & Richardson, 2007). The measures in their battery have shown relations to classroom learning behaviors (Denham, Warren-Khot, Bassett, Wyatt, & Perna, 2012) and concurrent and future academic achievement (Brock, Rimm-Kaufman, Nathanson, & Grimm, 2009) in young children.

The present work extends these efforts in several ways. First, it introduces a particular mix of easy to administer LRCSR measures especially suitable for use in early childhood education settings. Additionally, it broadens the scope of the data supporting the relations of these LRCSR measures to learning in those settings. As with the measures in the Chicago School Readiness Project battery, the measures we have identified are related to academic achievement and teacher reported LRCSR, relations that have not been demonstrated for many CSR measures appropriate for use with pre-K age children. In addition, however, this study has further explored relations with achievement gains, assessed developmental change over time, and demonstrated both test-retest and internal consistency reliability for the better performing measures that emerged from our comparative evaluation.

There are, of course, limitations to the research presented here. For practical reasons, it was not possible to collect data and conduct comparative analyses for all the CSR measures that have been used with pre-K age children. The selections we made may have omitted some measures that would have performed as well or better than those chosen. In particular, we would expect some of the computer-based measures to perform well on our criteria. Nonetheless, we excluded them to focus on measures that did not

require computer support or Internet connections, which we believe makes them more accessible and easily used in pre-K classroom settings, especially for potential use by teachers.

We also acknowledge the uncertain generalizability of the findings based on the sample of pre-K children that provided the data for this study. Though the initial sample included more than 500 children drawn from a relatively large number of schools and community childcare centers, it was of necessity limited to children whose parents consented to their participation. We have no data for the 40% of the children in those classrooms whose parents did not return consent forms (only a very few actively declined to consent) and thus have no basis for determining if they were systematically different from those consented in ways that might have affected our findings. And, though the sample was diverse with regard to gender, race, and economic status, it was fundamentally a convenience sample, not a probability sample of a defined population of pre-K children. Because of the span of schools, childcare centers, and classrooms, we have some confidence that this sample represented fairly typical pre-K age children in the middle Tennessee region, but no assurance that similarly constructed samples in other parts of the country would have produced comparable findings.

We also must emphasize that the fact that the LRCSR measures identified in this study are predictive of later achievement and achievement gains does not mean that they represent causal factors for those outcomes. Our purpose in this study was not to attempt to establish causal relations but, among other objectives, to identify measures that might be especially appropriate as outcome variables in research that does investigate causal influences. With conceptually relevant and responsive measures in hand, a key question for future research is what practical interventions or teacher practices are capable of increasing pre-K children's LRCSR skills. There is some evidence using one or another of the measures identified here that such effects are possible (e.g., Bierman et al., 2008; Fuhs, Farran, & Nesbitt, 2013; Raver et al., 2011), but also some less encouraging findings (e.g., Barnett et al., 2008).

Assuming that LRCSR can be boosted, an even more important question is whether doing so for pre-K children will, in turn, lead to greater learning and increased academic achievement. With regard to that question, we believe the concurrence between the LRCSR measures identified in this study and teacher ratings of LRCSR-related behaviors in the classroom is especially important. A very plausible theory of action for the potential effects on academic achievement of interventions that increase LRCSR is that they are mediated by the kinds of LRCSR-related behaviors teachers observe in the classrooms, for example, engagement in learning activities, persistence in completing tasks, attentiveness to teachers' instructions, and the like. Testing such causal and mediational relations is best done via randomized experiments and goes beyond the scope of the present study but is a promising area for further research aimed at improving the effectiveness of pre-K instruction. Based on the evidence developed in the present study, the well-performing LRCSR measures we have identified should be quite appropriate for supporting such research.

## References

- Allan, N. P., & Lonigan, C. J. (2011). Examining the dimensionality of effortful control in preschool children and its relation to academic and



- socioemotional indicators. *Developmental Psychology*, 47, 905–915. <http://dx.doi.org/10.1037/a0023748>
- Baddeley, A., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). New York, NY: Academic Press.
- Barnett, W. S., Jung, K., Yarosz, D. J., Thomas, J., Hornbeck, A., Stechuk, R., & Burns, S. (2008). Educational effects of the Tools of the Mind curriculum: A randomized trial. *Early Childhood Research Quarterly*, 23, 299–313. <http://dx.doi.org/10.1016/j.ecresq.2008.03.001>
- Beck, D. M., Schaefer, C., Pang, K., & Carlson, S. M. (2011). Executive function in preschool children: Test–retest reliability. *Journal of Cognition and Development*, 12, 169–193. <http://dx.doi.org/10.1080/15248372.2011.563485>
- Beck, L. H., Bransome, E. D., Jr., Mirsky, A. F., Rosvold, H. E., & Sarason, I. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology*, 20, 343–350. <http://dx.doi.org/10.1037/h0043220>
- Bender, L. (1938). *A visual motor Gestalt test and its clinical use* (Research Monograph No. 3). New York, NY: American Orthopsychiatric Association.
- Berch, D. B., Krikorian, R., & Huha, E. M. (1998). The Corsi Block-Tapping Task: Methodological and theoretical considerations. *Brain and Cognition*, 38, 317–338. <http://dx.doi.org/10.1006/brcg.1998.1039>
- Bierman, K. L., Nix, R. L., Greenberg, M. T., Blair, C., & Domitrovich, C. E. (2008). Executive functions and school readiness intervention: Impact, moderation, and mediation in the Head Start REDI program. *Development and Psychopathology*, 20, 821–843. <http://dx.doi.org/10.1017/S0954579408000394>
- Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, 78, 647–663. <http://dx.doi.org/10.1111/j.1467-8624.2007.01019.x>
- Blair, C., & Willoughby, M. T. (2006a). *Measuring executive function in young children: Item Selection*. Chapel Hill, NC: Pennsylvania State University and University of North Carolina.
- Blair, C., & Willoughby, M. T. (2006b). *Measuring executive function in young children: Operation Span*. Chapel Hill, NC: Pennsylvania State University and University of North Carolina.
- Blair, C., & Willoughby, M. T. (2006c). *Measuring executive function in young children: The Pig Game*. Chapel Hill, NC: Pennsylvania State University and University of North Carolina.
- Blair, C., & Willoughby, M. T. (2006d). *Measuring executive function in young children: Silly Sounds Game*. Chapel Hill, NC: Pennsylvania State University and University of North Carolina.
- Blair, C., & Willoughby, M. T. (2006e). *Measuring executive function in young children: Spatial Conflict*. Chapel Hill, NC: Pennsylvania State University and University of North Carolina.
- Blair, C. B., & Willoughby, M. T. (2006f). *Measuring executive function in young children: Spatial Conflict II: Arrows*. Chapel Hill, NC: The Pennsylvania State University and the University of North Carolina at Chapel Hill.
- Bodovski, K., & Farkas, G. (2007). Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *The Elementary School Journal*, 108, 115–130. <http://dx.doi.org/10.1086/525550>
- Brock, L. L., Rimm-Kaufman, S. E., Nathanson, L., & Grimm, K. J. (2009). The contributions of ‘hot’ and ‘cool’ executive function to children’s academic achievement, learning-related behaviors, and engagement in kindergarten. *Early Childhood Research Quarterly*, 24, 337–349. <http://dx.doi.org/10.1016/j.ecresq.2009.06.001>
- Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology*, 28, 595–616. [http://dx.doi.org/10.1207/s15326942dn2802\\_3](http://dx.doi.org/10.1207/s15326942dn2802_3)
- Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children’s theory of mind. *Child Development*, 72, 1032–1053. <http://dx.doi.org/10.1111/1467-8624.00333>
- Cooper, D. H., & Farran, D. C. (1988). Behavioral risk factors in kindergarten. *Early Childhood Research Quarterly*, 3, 1–19. [http://dx.doi.org/10.1016/0885-2006\(88\)90026-9](http://dx.doi.org/10.1016/0885-2006(88)90026-9)
- Cooper, D. H., & Farran, D. C. (1991). *The Cooper-Farran Behavioral Rating Scales*. Brandon, VT: Clinical Psychology Publishing Co., Inc.
- Cronbach, L. J., & Furby, L. (1970). How we should measure “change”—or should we? *Psychological Bulletin*, 74, 68–80. <http://dx.doi.org/10.1037/h0029382>
- Davie, R., Butler, H., & Goldstein, H. (1972). *From birth to seven: The second report of the National Child Development Study* (1958 Cohort). London, UK: Longman.
- Davis, H. L., & Pratt, C. (1995). The development of children’s theory of mind: The working memory explanation. *Australian Journal of Psychology*, 47, 25–31. <http://dx.doi.org/10.1080/00049539508258765>
- Davoudzadeh, P., McTernan, M., & Grimm, K. (2015). Early school readiness predictors of grade retention from kindergarten through eighth grade: A multilevel discrete-time survival analysis approach. *Early Childhood Research Quarterly*, 32, 183–192. <http://dx.doi.org/10.1016/j.ecresq.2015.04.005>
- Denham, S. A., Warren-Khot, H. K., Bassett, H. H., Wyatt, T., & Perna, A. (2012). Factor structure of self-regulation in preschoolers: Testing models of a field-based assessment for predicting early school readiness. *Journal of Experimental Child Psychology*, 111, 386–404. <http://dx.doi.org/10.1016/j.jecp.2011.10.002>
- Diamond, A. (1990). Developmental time course in human infants and infant monkeys, and the neural bases of, inhibitory control in reaching. *Annals of the New York Academy of Sciences*, 608, 637–676. <http://dx.doi.org/10.1111/j.1749-6632.1990.tb48913.x>
- Diamond, A., Carlson, S. M., & Beck, D. M. (2005). Preschool children’s performance in task switching on the dimensional change card sort task: Separating the dimensions aids the ability to switch. *Developmental Neuropsychology*, 28, 689–729. [http://dx.doi.org/10.1207/s15326942dn2802\\_7](http://dx.doi.org/10.1207/s15326942dn2802_7)
- Diamond, A., Prevor, M. B., Callender, G., & Druin, D. P. (1997). Prefrontal cognitive deficits in children treated early and continuously for PKU. *Monographs of the Society for Research in Child Development*, 62(4), Serial No.252.
- Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the ability to remember what I said and to “do as I say, not as I do.” *Developmental Psychobiology*, 29, 315–334.
- Dibner, A. S., & Korn, E. J. (1969). Group administration of the Bender-Gestalt test to predict early school performance. *Journal of Clinical Psychology*, 25, 263–268. [http://dx.doi.org/10.1002/1097-4679\(196907\)25:3<263::AID-JCLP2270250311>3.0.CO;2-D](http://dx.doi.org/10.1002/1097-4679(196907)25:3<263::AID-JCLP2270250311>3.0.CO;2-D)
- Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, 43, 1428–1446. <http://dx.doi.org/10.1037/0012-1649.43.6.1428>
- Duncan, S. E., & DeAvila, E. A. (1985). *How to administer Pre-LAS*. Monterey, CA: CTB/McGraw-Hill.
- Fuhs, M. W., Farran, D. C., & Nesbitt, K. T. (2013). Prekindergarten children’s executive functioning skills and achievement gains: The utility of direct assessments and teacher ratings. *School Psychology Quarterly*, 28, 347–359. <http://dx.doi.org/10.1037/spq0000031>
- Fuhs, M. W., Farran, D. C., & Nesbitt, K. T. (2015). Prekindergarten children’s executive function skills and achievement gains: Comparing direct assessments and teacher ratings. *Journal of Educational Psychology*, 107, 207–221. <http://dx.doi.org/10.1037/a0037366>
- Fuhs, M. W., Nesbitt, K. T., Farran, D. C., & Dong, N. (2014). Longitudinal associations between executive functioning and academic skills



- across content areas. *Developmental Psychology*, 50, 1698–1709. <http://dx.doi.org/10.1037/a0036633>
- Gathercole, S., Brown, L., & Pickering, S. (2003). Working memory assessments at school entry as longitudinal predictors of National Curriculum attainment levels. *Educational and Child Psychology*, 20, 109–122.
- Gerardi-Caulton, G. (2000). Sensitivity to spatial conflict and the development of self-regulation in children 24–36 months of age. *Developmental Science*, 3, 397–404. <http://dx.doi.org/10.1111/1467-7687.00134>
- Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 3 1/2–7 years old on a Stroop-like day-night test. *Cognition*, 53, 129–153. [http://dx.doi.org/10.1016/0010-0277\(94\)90068-X](http://dx.doi.org/10.1016/0010-0277(94)90068-X)
- Graziano, P. A., Slavec, J., Ros, R., Garb, L., Hart, K., & Garcia, A. (2015). Self-regulation assessment among preschoolers with externalizing behavior problems. *Psychological Assessment*, 27, 1337–1348. <http://dx.doi.org/10.1037/pas0000113>
- Howse, R. B., Lange, G., Farran, D. C., & Boyles, C. D. (2003). Motivation and self-regulation as predictors of achievement in economically disadvantaged young children. *Journal of Experimental Education*, 71, 151–174. <http://dx.doi.org/10.1080/00220970309602061>
- Hughes, C. (1998). Finding your marbles: Does preschoolers' strategic behavior predict later understanding of mind? *Developmental Psychology*, 34, 1326–1339. <http://dx.doi.org/10.1037/0012-1649.34.6.1326>
- Hughes, C., & Ensor, R. (2011). Individual differences in growth in executive function across the transition to school predict externalizing and internalizing behaviors and self-perceived academic success at 6 years of age. *Journal of Experimental Child Psychology*, 108, 663–676. <http://dx.doi.org/10.1016/j.jecp.2010.06.005>
- Jacob, R., & Parkinson, J. (2015). The potential for school-based interventions that target executive function to improve academic achievement: A review. *Review of Educational Research*, 85, 512–552. <http://dx.doi.org/10.3102/0034654314561338>
- Jacques, S., & Zelazo, P. D. (2001). The Flexible Item Selection Task (FIST): A measure of executive function in preschoolers. *Developmental Neuropsychology*, 20, 573–591. [http://dx.doi.org/10.1207/S15326942DN2003\\_2](http://dx.doi.org/10.1207/S15326942DN2003_2)
- Jones, L. B., Rothbart, M. K., & Posner, M. I. (2003). Development of executive attention in preschool children. *Developmental Science*, 6, 498–504. <http://dx.doi.org/10.1111/1467-7687.00307>
- Kagan, J., Rosman, B. L., Day, D., Albert, J., & Phillips, W. (1964). Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monographs*, 78(1), No. 578.
- Katz, L. G. (1993). Dispositions: Definitions and implications for early childhood practices. *Perspectives From EECE: A Monograph Series*. Urbana, IL: ERIC Clearinghouse on Elementary and Early Childhood Education.
- Katz, L. (2002). "Not all dispositions are desirable": Implications for assessment. *Assessment in Education: Principles, Policy & Practice*, 9, 53–54. <http://dx.doi.org/10.1080/09695940220119175>
- Kochanska, G., Murray, K. T., & Harlan, E. T. (2000). Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology*, 36, 220–232. <http://dx.doi.org/10.1037/0012-1649.36.2.220>
- Kochanska, G., Murray, K., Jacques, T. Y., Koenig, A. L., & Vandegeest, K. A. (1996). Inhibitory control in young children and its role in emerging internalization. *Child Development*, 67, 490–507. <http://dx.doi.org/10.2307/1131828>
- Koppitz, E. M. (1973). Bender Gestalt Test performance and school achievement: A 9-year study. *Psychology in the Schools*, 10, 280–284. [http://dx.doi.org/10.1002/1520-6807\(197307\)10:3<280::A1D-PITS2310100302>3.0.CO;2-6](http://dx.doi.org/10.1002/1520-6807(197307)10:3<280::A1D-PITS2310100302>3.0.CO;2-6)
- Kirkorian, R., Bartok, J., & Gay, N. (1994). Tower of London procedure: A standard method and developmental data. *Journal of Clinical and Experimental Neuropsychology*, 16, 840–850. <http://dx.doi.org/10.1080/01688639408402697>
- Kuhl, J., & Kraska, K. (1993). Self-regulation: Psychometric properties of a computer-aided instrument. *German Journal of Psychology*, 17, 11–24.
- Lan, X., Legare, C. H., Ponitz, C. C., Li, S., & Morrison, F. J. (2011). Investigating the links between the subcomponents of executive function and academic achievement: A cross-cultural analysis of Chinese and American preschoolers. *Journal of Experimental Child Psychology*, 108, 677–692. <http://dx.doi.org/10.1016/j.jecp.2010.11.001>
- Li-Grining, C. P., Votruba-Drzal, E., Maldonado-Carreño, C., & Haas, K. (2010). Children's early approaches to learning and academic trajectories through fifth grade. *Developmental Psychology*, 46, 1062–1077. <http://dx.doi.org/10.1037/a0020066>
- Luria, A. R., Pribram, K. H., & Homskaya, E. D. (1964). An experimental analysis of the behavior disturbance produced by a left frontal arachnoidal endothelioma (meningioma). *Neuropsychologia*, 2, 257–280. [http://dx.doi.org/10.1016/0028-3932\(64\)90034-X](http://dx.doi.org/10.1016/0028-3932(64)90034-X)
- MacCoby, E., Dowley, E., Hagen, J., & Degerman, R. (1965). Activity level and intellectual functioning in normal preschool children. *Child Development*, 36, 761–770. <http://dx.doi.org/10.2307/1126921>
- Martin, R. P. (1988). *The Temperament Assessment Battery for Children*. Brandon, VT: Clinical Psychology Publishing.
- Matthews, J., Ponitz, C. C., & Morrison, F. J. (2009). Early gender differences in self-regulation and academic achievement. *Journal of Educational Psychology*, 101, 689–704. <http://dx.doi.org/10.1037/a0014240>
- McClelland, M. M., Cameron, C. E., Connor, C. M., Farris, C. L., Jewkes, A. M., & Morrison, F. J. (2007). Links between behavioral regulation and preschoolers' literacy, vocabulary, and math skills. *Developmental Psychology*, 43, 947–959. <http://dx.doi.org/10.1037/0012-1649.43.4.947>
- McClelland, M. M., Morrison, F. J., & Holmes, D. L. (2000). Children at risk for early academic problems: The role of learning-related social skills. *Early Childhood Research Quarterly*, 15, 307–329. [http://dx.doi.org/10.1016/S0885-2006\(00\)00069-7](http://dx.doi.org/10.1016/S0885-2006(00)00069-7)
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, 41, 49–100. <http://dx.doi.org/10.1006/cogp.1999.0734>
- Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind? *Journal of Learning Disabilities*, 44, 472–488. <http://dx.doi.org/10.1177/0022219411414010>
- Osborn, A. F., Butler, N. R., & Morris, A. C. (1984). *The social life of Britain's five year olds: A report of the Child Health and Education Study*. London, UK: Routledge and Kegan Paul.
- Petrides, M., & Milner, B. (1982). Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*, 20, 249–262. [http://dx.doi.org/10.1016/0028-3932\(82\)90100-2](http://dx.doi.org/10.1016/0028-3932(82)90100-2)
- Pickering, S. J., & Gathercole, S. E. (2001). *Working memory test battery for children*. London: Psychological Corporation.
- Ponitz, C. C., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral regulation and its contribution to kindergarten outcomes. *Developmental Psychology*, 45, 605–619. <http://dx.doi.org/10.1037/a0015365>
- Posner, M. I., & Rothbart, M. K. (2000). Developing mechanisms of self-regulation. *Development and Psychopathology*, 12, 427–441. <http://dx.doi.org/10.1017/S0954579400003096>
- Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's Impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development*, 82, 362–378. <http://dx.doi.org/10.1111/j.1467-8624.2010.01561.x>
- Reed, M. A., Pien, D. L., & Rothbart, M. K. (1984). Inhibitory self-control in preschool children. *Merrill-Palmer Quarterly*, 30, 131–147.

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology, 45*, 958–972. <http://dx.doi.org/10.1037/a0015861>

Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome, E. D., & Beck, L. H. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology, 20*, 343–350. <http://dx.doi.org/10.1037/h0043220>

Rothbart, M. K., & Ahadi, S. A. (1994). Temperament and the development of personality. *Journal of Abnormal Psychology, 103*, 55–66. <http://dx.doi.org/10.1037/0021-843X.103.1.55>

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child Development, 72*, 1394–1408. <http://dx.doi.org/10.1111/1467-8624.00355>

Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., & Posner, M. I. (2004). Development of attentional networks in childhood. *Neuropsychologia, 42*, 1029–1040. <http://dx.doi.org/10.1016/j.neuropsychologia.2003.12.012>

Schmitt, S., Pratt, M., & McClelland, M. (2014). Examining the validity of behavioral self-regulation tools in predicting preschoolers' academic achievement. *Early Education and Development, 25*, 641–660. <http://dx.doi.org/10.1080/10409289.2014.850397>

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-Regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly, 22*, 173–187. <http://dx.doi.org/10.1016/j.ecresq.2007.01.002>

Strommen, E. A. (1973). Verbal self-regulation in a children's game: Impulsive errors on "Simon says." *Child Development, 44*, 849–853. <http://dx.doi.org/10.2307/1127737>

Ward, H., Shum, D., McKinlay, L., Baker-Tweney, S., & Wallace, G. (2005). Development of prospective memory: Tasks based on the prefrontal-lobe model. *Child Neuropsychology, 11*, 527–549. <http://dx.doi.org/10.1080/09297040490920186>

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth Edition* (WISC-IV). San Antonio, TX: The Psychological Corporation.

Weintraub, S., Bauer, P. J., Zelazo, P. D., Wallner-Allen, K., Dikmen, S. S., Heaton, R. K., . . . Gershon, R. C. (2013). I. NIH Toolbox Cognition Battery (CB): Introduction and pediatric data. *Monographs of the Society for Research in Child Development, 78*, 1–15. <http://dx.doi.org/10.1111/mono.12031>

Welsh, J. A., Nix, R. L., Blair, C., Bierman, K. L., & Nelson, K. E. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of Educational Psychology, 102*, 43–53. <http://dx.doi.org/10.1037/a0016738>

Willoughby, M. T., & Blair, C. B., & the Family Life Project Investigators. (2016). Measuring executive function in early childhood: A case for formative measurement. *Psychological Assessment, 28*, 319–330. <http://dx.doi.org/10.1037/pas0000152>

Willoughby, M. T., Blair, C. B., Wirth, R. J., & Greenberg, M. (2012). The measurement of executive function at age 5: Psychometric properties and relationship to academic achievement. *Psychological Assessment, 24*, 226–239. <http://dx.doi.org/10.1037/a0025361>

Willoughby, M. T., Wirth, R. J., & Blair, C. B. (2011). Contributions of modern measurement theory to measuring executive function in early childhood: An empirical demonstration. *Journal of Experimental Child Psychology, 108*, 414–435. <http://dx.doi.org/10.1016/j.jecp.2010.04.007>

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Rolling Meadows, IL: Riverside Publishing.

Wright, J. C. (1971). *Kansas Reflection-Impulsivity Scale for Preschoolers (KRISP)*. St. Louis, MO: CEMREL, Inc.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols, 1*, 297–301. <http://dx.doi.org/10.1038/nprot.2006.46>

Zelazo, P. D., & Bauer, P. J. (2013). National Institute of Health Toolbox Cognition Battery (NIH Toolbox CB): Validation for children between 3 and 15 years. *Monographs of the Society for Research in Child Development, 78*, 1–172.

Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development, 11*, 37–63. [http://dx.doi.org/10.1016/S0885-2014\(96\)90027-1](http://dx.doi.org/10.1016/S0885-2014(96)90027-1)

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*, 3–17. <http://dx.doi.org/10.1207/s15326985ep25012>

Appendix A

Candidate Direct Child Assessment Tasks Categorized by the Most Salient CSR Component Skill

Instrument	Task
Sustained attention—attending to and sustaining focus on a task	
1. Copying tasks	
Bender Gestalt Test (Bender, 1938; Dibner & Korn, 1969; Koppitz, 1973)	Children copy nine simple geometric designs exactly.
Copy Design (Davie et al., 1972; Osborn et al., 1984)	Children copy eight simple geometric designs exactly.
2. Matching tasks	
Matching Familiar Figures Test (Kagan et al., 1964)	Children select a picture that matches a target picture; accuracy is a measure of attention, over-fast reaction times assess impulsivity.

(Appendices continue)



## Appendix A (continued)

Instrument	Task
Kansas Reflection-Impulsivity Scale for Preschoolers (Wright, 1971)	Children select a picture that matches a target picture; accuracy is a measure of attention, over-fast reaction times assess impulsivity.
Tower of London Task (Kirkorian et al., 1994; Ward et al., 2005)	Children build a tower to match a picture using blocks and a peg for stacking.
3. Stimulus-response tasks	
Continuous Performance Test (Beck et al., 1956)	Children press a button when a computer-generated target stimulus appears and inhibit responses to non-target stimuli. Correct responses measure of sustained attention; incorrect responses measure impulsivity.
Self-Regulation Test for Children (Howse et al., 2003; Kuhl & Kraska, 1993)	Children press the button that matches the target on a screen; can involve distractors to make the task more difficult.
Attention shifting—shifting focus within or between tasks as situations demand	
Something's the Same/Item Selection (Blair & Willoughby, 2006a)	Children categorize colored pictures first by the object, then switch to sort by color.
Dimensional Change Card Sort (Diamond et al., 2005; Zelazo, 2006) & variants (Wisconsin Card Sorting Task)	Children sort a set of cards by shape, then switch to sort by color. A more difficult version adds a variable cue that indicates the sorting rule.
Flexible Item Selection Task (Jacques & Zelazo, 2001)	Children select a pair of cards that match on one dimension (e.g., shape, color), then must select a different pair that matches on another dimension.
Working memory—active maintenance and manipulation of information in memory	
Operation Span (Blair & Willoughby, 2006b)	Children must recall a series of objects shown inside a picture of a house. A color distractor adds difficulty.
Self-Ordered Pointing (Petrides & Milner, 1982).	Children are presented a set of pages divided into four sections; they must go through and point to a different remembered picture on each page.
Backward Digit Span (Davis & Pratt, 1996; Pickering & Gathercole, 2001).	Children recall a series of orally presented digits backwards.
Corsi Block Tapping (Berch et al., 1998)	Children reproduce the sequence in which the assessor taps a series of blocks with sequences of increasing length.
Inhibitory control—volitional inhibition of a prepotent response to complete a task	
1. Stroop-like tasks	
Silly Sounds Game (Blair & Willoughby, 2006d)	Children meow to pictures of dogs and bark to pictures of cats.
Day/Night (Carlson & Moses, 2001; Gerstadt et al., 1994)	Children say "night" to sun pictures and "day" to moon pictures.
Grass/Snow (Carlson & Moses, 2001)	Children say "green" to snow pictures, and "white" to grass pictures.
2. Stroop-like tasks with motor response	
Bear & Dragon and variants (Jones et al., 2003; Reed et al., 1984)	A bear puppet and dragon puppet give children tasks (touching feet, hopping, etc.); then, they are asked to only perform tasks given by the bear puppet, not those given by the dragon puppet.
Simon Says (Carlson, 2005; Strommen, 1973)	Children perform certain tasks (touching their feet, hopping, etc.) only when the assessor precedes the command with "Simon Says."
Head-to-Toes; Head-Toes-Knees-Shoulders (Ponitz et al., 2009; McClelland et al., 2007)	Children do the opposite of what the assessor requests; e.g., if asked to touch their head, they touch their toes.
Luria Hand Game (Hughes, 1998; Luria et al., 1964)	Similar to the Head-to-Toes task, children do the opposite of what the assessor indicates using hand signals (e.g., holding up a fist vs. one finger).
Peg- or Finger-Tapping (Diamond & Taylor, 1996; Diamond et al., 1997; Smith-Donald et al., 2007)	Children tap a peg, pencil, or finger twice when the experimenter taps once, and vice versa.
Pig Game (Blair & Willoughby, 2006c)	In a series of animal pictures, children press a button when they see animals that aren't pigs, and don't press the button when they do see pigs.
3. Spatial Conflict/Simon Tasks	
Spatial conflict (Blair & Willoughby, 2006e)	Target pictures are presented on the left side of a paper and children point to the target pictures with their right hand, and vice versa.

(Appendices continue)

Appendix A (continued)

Instrument	Task
Spatial conflict (Gerardi-Caulton, 2000)	Computerized version of the task above; children push a key on one side of the keyboard for target on the opposite side of the computer screen.
4. Flanker Tasks	
Attention Network Task/Flanker Task (Ponitz et al., 2009; Rueda et al., 2004)	Children indicate the direction of a target flanked by same/opposite direction distractors; e.g., feed the central fish by pressing a button corresponding to the direction which the middle fish is swimming when flanked by fish swimming the same or opposite direction.
Effortful control—suppression of impulsive or premature responses when required by a task	
Snack delay; gift delay (Kochanska et al., 2000)	Variety of delay tasks in which children must wait before eating a cookie, open a gift, etc.
Tower: Turn-taking (Kochanska et al., 1996)	Assessor and child take turns placing blocks on a tower; children must wait their turn without reminders.
Whisper (Kochanska, et al., 1996)	Children see pictures of familiar cartoon characters and whisper their names; number whispered vs. shouted or said in normal voice is scored.
Walk-a-Line Slowly/Draw-a-Line Slowly (Maccoby et al., 1965)	Children walk or draw a line at normal speed, then do the same thing slowly; time difference between regular and slow trials is scored.
Turtle and Rabbit (Kochanska et al., 1996)	Children are given “fast” rabbit and a “slow” turtle toys and move them along a path; scored for accuracy in negotiating the path and the time difference between fast and slow trials.

Appendix B

Scoring Scheme for the Child Measures of Learning-Related Cognitive Self-Regulation

Rescaled score	Scores in the original metric for each measure					
	Peg Tapping	HTKS	KRISP	DCCS	Copy Design	Backwards Digit Span
0	≤5	≤7	≤25	0	0	0
1	6–7	8–15	26–29	1	1	1
2	8–9	16–23	30–32	1	2	2
3	10–12	24–31	33–36	2	3	3
4	13–14	32–38	37–39	2	4–5	4
5	>14	>38	>39	3	>5	≥5

*Note.* HTKS = Head-Toes-Knees-Shoulders; KRISP = Kansas Reflection-Impulsivity Scale for Preschoolers; DCCS = Dimensional Change Card Sort. The six learning-related cognitive self-regulation (LRCSR) measures identified in this paper were scored on different scales (e.g., 0–3 for DCCS, 0–52 for HTKS), complicating the construction of a total score for all six measures together. One solution is to rescale the scores on each measure to a common scale, then sum them for a total score. We found that a 0–5 point scale format worked well for this purpose. To determine which original scores should be rescored into each value on this common scale, we took advantage of the linear relation between children’s age and their scores on each measure. Using data from the initial sample, we regressed the scores for each measure on age and used the results to estimate the scores in the original metric expected at ages 4.0, 4.5, 5.0, 5.5, and 6.0, spanning the pre-K age range. These estimates were then used as break points for rescaling each original score into the 0–5 format. The resulting procedure is shown above. In the initial sample with which this scheme was constructed, correlations between rescaled scores and those in the original metric ranged from .92 to 1.00 across measures and the Time 1 (beginning of pre-K) and Time 2 (end of pre-K) measurement waves. They also performed well for the Time 3 end of kindergarten measures with correlations from .82 to .98. When applied to the Time 1 and 2 data from the cross-validation sample, the correlations ranged from .91 to 1.00. The total scores produced by summing the rescaled scores across all six items showed correlations from .94 to .99 with the factor scores for Time 1, 2, and 3 in the initial sample, and correlations from .97 to .99 with the Time 1 and 2 factor scores in the cross-validation sample.

Received May 7, 2014  
Revision received February 2, 2017  
Accepted February 21, 2017 ■



# Effects of a Year Long Supplemental Reading Intervention for Students With Reading Difficulties in Fourth Grade

Jeanne Wanzek  
Vanderbilt University

Yaacov Petscher  
Florida State University

Stephanie Al Otaiba, Brenna K. Rivas, and  
Francesca G. Jones  
Southern Methodist University

Shawn C. Kent  
University of Houston

Christopher Schatschneider  
Florida State University

Paras Mehta  
University of Houston

Research examining effective reading interventions for students with reading difficulties in the upper elementary grades is limited relative to the information available for the early elementary grades. In the current study, we examined the effects of a multicomponent reading intervention for students with reading comprehension difficulties. We used a partially nested analysis with latent variables to adequately match the design of the study and provide the necessary precision of intervention effects. We examined the effects of the intervention on students' latent word reading, latent vocabulary, and latent reading comprehension. In addition, we examined whether these effects differed for students of varying levels of reading or English language proficiency. Findings indicated the treatment significantly outperformed the comparison on reading comprehension (Effect Size = 0.38), but no overall group differences were noted on word reading or vocabulary. Students' initial word reading scores moderated this effect. Reading comprehension effects were similar for English learner and non-English learner students.

## *Educational Impact and Implications Statement*

This study examined the effects of a multi-component reading intervention for students with reading difficulties in fourth grade. Findings indicated students receiving the intervention made greater gains in reading comprehension than students who did not receive the intervention. This finding was similar for students who were English learners or non-English learners. However, students with higher initial word reading scores benefited more from the intervention. These findings suggest students receiving the intervention made progress in closing the gap between their current level of performance and expected levels of performance in reading comprehension.

**Keywords:** reading intervention, reading difficulties, elementary

Students with reading difficulties can benefit from supplemental reading instruction provided in small groups; reading interventions at the elementary level have demonstrated power for preventing and remediating many reading difficulties (Blachman et al., 2004;

Mathes et al., 2005; O'Connor, Fulmer, Harty, & Bell, 2005; Torgesen et al., 1999; Vellutino et al., 1996). However, research examining effective reading interventions for students with reading difficulties in the upper elementary grades is limited relative to

This article was published Online First March 27, 2017.

Jeanne Wanzek, Department of Special Education, Vanderbilt University; Yaacov Petscher, Florida Center for Reading Research, Florida State University; Stephanie Al Otaiba, Brenna K. Rivas, and Francesca G. Jones, Simmons School of Education Southern Methodist University; Shawn C. Kent, Department of Educational Leadership and Policy Studies, University of Houston; Christopher Schatschneider, Florida Center for Reading Research and Department of Psychology, Florida State University; Paras Mehta, Department of Psychology, University of Houston.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A150269 to Vanderbilt University. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Jeanne Wanzek, Department of Special Education, Vanderbilt University, 110 Magnolia Circle, Nashville, TN 37203. E-mail: jeanne.wanzek@vanderbilt.edu

the information available for the early elementary grades (Wanzek, Wexler, Vaughn, & Ciullo, 2010). The need for effective reading interventions for students with reading difficulties in the upper elementary grades is essential given the large numbers of students who continue to struggle with reading at these grade levels (National Center for Educational Statistics, 2016).

### Reading Interventions for Upper Elementary Students

The research available on reading interventions related to upper elementary students with reading difficulties demonstrates positive effects for interventions providing instruction in comprehension or word recognition (Wanzek et al., 2010). Higher effects were noted for interventions related specifically to comprehension instruction. For example, large mean effects across comprehension measures were noted in two experimental studies of comprehension strategy instruction for students with reading difficulties (Mason, 2004; Miranda et al., 1997). However, the upper elementary research, including these comprehension interventions, has also largely examined intervention effects on proximal, researcher-developed measures. In fact, 15 of the 24 studies synthesized by Wanzek et al. (2010) used only researcher-developed measures. Researcher-developed measures often result in higher effects than standardized measures of the same constructs (Scammacca et al., 2007; Swanson, Hoskyn, & Lee, 1999). Thus, the lack of information on the effects of providing comprehension interventions on standardized measures represents a gap in the knowledge base on upper elementary reading interventions.

Additionally, Wanzek et al. (2010) reported that most of research thus far on upper elementary reading interventions for students with reading difficulties has been conducted with relatively brief interventions (e.g., 15-min sessions; less than 6 weeks) that examined single instructional strategies (e.g., main idea strategy). These studies provide important information regarding effective practices that could be incorporated in reading interventions to accelerate student learning. Knowledge of student outcomes when effective practices for various reading components are put together to form more comprehensive interventions for struggling readers is also needed.

In fact, some of the highest effects in the upper elementary reading intervention literature have come from multicomponent interventions (Wanzek et al., 2010). Though there are only a few of these studies in the literature (e.g., O'Connor et al., 2002; Ritchey, Silverman, Montanaro, Speece, & Schatschneider, 2012; Therrien, Wickstrom, & Jones, 2006; Vadasy & Sanders, 2008; Wanzek & Roberts, 2012), the findings suggest the possible importance of addressing multiple reading components in reading intervention for these older students. Three of these studies demonstrated moderate to large, significant effects on norm-referenced measures of comprehension or broad reading achievement (O'Connor et al., 2002; Therrien et al., 2006; Vadasy & Sanders, 2008). The effect sizes ranged from 0.37 to 1.87. The interventions in these studies included instruction in reading comprehension along with additional instruction in word reading (O'Connor et al., 2002), fluency (O'Connor et al., 2002; Therrien et al., 2006; Vadasy & Sanders, 2008), and/or vocabulary (Vadasy & Sanders, 2008). The findings suggest students with reading difficulties at the upper elementary level may benefit most when interventions focus on multiple elements of reading, providing opportunities for

students to integrate reading practices to read and understand text. In an earlier synthesis of interventions for students with learning disabilities, Swanson et al. (1999) reported the highest effects for interventions that combine direct instruction of content with strategy instruction. Most of the multiple component reading interventions conducted at the upper elementary level have incorporated both types of instruction. Several other syntheses for older students confirm the value of multicomponent interventions (Kamil et al., 2008; Scammacca et al., 2007; Torgesen et al., 2007).

The previous research also suggests some differential effects for English learners (ELs) with reading difficulties relative to their non-EL peers (Kieffer, 2008). In particular, ELs are at a markedly greater risk of late-emerging (after Grade 3) reading difficulties (Kieffer, 2010, 2014), suggesting reading foundation skills such as word reading may be mastered more easily. But, many ELs may struggle later with understanding texts that have more complex syntax, vocabulary, or background knowledge needs. Previous fourth grade interventions have noted higher effects for ELs in reading intervention on word reading measures but not on comprehension or vocabulary measures (Wanzek & Roberts, 2012). Thus, examining the differential effects of ELs with a multicomponent, comprehension focused reading intervention program could provide additional evidence regarding for whom a reading intervention is most valuable.

### Passport to Literacy

One multicomponent reading intervention that is widely used in schools across the United States is Passport to Literacy. Passport to Literacy is a packaged program that applies principles of behavioral learning theory and cognitive psychology (Flavell, 1992; Palincsar & Brown, 1984), providing explicit instruction and strategies for reasoning in the foundational skills of reading (e.g., decoding, word reading) as well as reading comprehension and vocabulary. Semiscripted lessons are built sequentially to help students acquire missing foundational reading skills, increase background knowledge, and build strategies for comprehending text.

Although Passport to Literacy is widely used, there is a lack of independent research on the program's effectiveness. We conducted one initial study of the Passport to Literacy intervention with fourth grade students. This study was the first causal study conducted on Passport to Literacy and also the first to examine outcomes on standardized measures of reading achievement. Fourth grade students scoring below the 30th percentile in reading comprehension ( $n = 221$ ) were randomly assigned to receive the standard implementation of the Passport to Literacy intervention or typical school services. The intervention was provided in small groups of four to seven students for 30 min, 4 days a week throughout the school year ( $M = 90.45$  lessons). There were no effects for Passport to Literacy on standardized measures of word reading or fluency, but small effects were noted on standardized measures of reading comprehension (Effect Size (ES) = 0.14 to 0.28). Exploratory analyses indicated the intervention effects differed by students' comprehension abilities. Students' exhibiting low levels of comprehension demonstrated no increased benefit of the Passport to Literacy standard intervention. In other words, the multicomponent Passport to Literacy intervention demonstrated



average increased outcomes on reading comprehension, but was least effective for students with the lowest comprehension levels.

In the current study, we build upon this previous study to examine the effects of Passport to Literacy with a larger sample. This larger sample allows for a more sophisticated analysis that matches the design of the study taking into account the differing clustering structures of the treatment and comparison groups. In addition, the larger sample allows us to be more precise in measuring student reading achievement through the use of latent variables. By using latent variables, the impact and exploratory analyses reflect a stronger test of theory as effects are less due to assessment-specific outcomes and more to the theoretical overlap among them. Finally, the larger sample included a large enough sample of ELs to examine other possible associations that may explain the differential effects noted in the first study.

### Study Purpose

The purpose of this study was to examine the effects of the standard implementation of the Passport to Literacy intervention for students with reading comprehension difficulties. We sought to examine the effects of this multicomponent intervention on students' word reading, vocabulary, and reading comprehension. In addition, we examined whether these effects differed for students with varying levels of reading or English language proficiency. Specifically, we examined the following:

1. What are the effects of Passport to Literacy on students' word reading, vocabulary, and reading comprehension?
2. Do these effects differ by initial reading achievement or English language level?

On the basis of the previous study of the intervention, we hypothesized that students with reading difficulties receiving the Passport to Literacy intervention would outperform students receiving typical school services in reading comprehension and not in word reading or vocabulary. We also hypothesized that students with higher initial levels of reading achievement on word reading, fluency, or comprehension would benefit more from the intervention. On the basis of previous reading intervention work for ELs we hypothesized more benefits of the multicomponent intervention for ELs on word reading outcomes than for their non-EL peers.

### Method

#### Participants

Four hundred fifty-one Grade 4 students who scored at or below the 30th percentile on the reading comprehension subtest of the Gates-MacGinitie Reading Tests (GMRT; MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2006) were selected for the study. The students came from 16 public elementary schools located across six school districts in three states. One school district was located in a large, urban metropolitan area; one district was located in a midsize city; and four districts were located in rural areas. Male students made up 49% of the sample. With regard to ethnicity, 46% of the students were identified as Hispanic. Of those who

reported language status, 13.2% of the total sample was flagged as having a primary language other than English or as currently receiving EL services. All schools provided only instruction in English. The racial composition of the sample was 35% Black, 44% White, 17% American Indian, 1% Asian, and 2% multiracial. Eighty-five percent of the students qualified for low income or free or reduced lunch programs. Fifteen percent were identified as having a disability. The majority of students with a disability were identified with a learning disability or a speech/language disability. There were no differences in any of the demographics between the two study groups.

A total of 40 students (9% of total sample) withdrew from their respective schools after the screening test. Attrition was 12% ( $n = 27$ ) in the treatment group and 6% ( $n = 13$ ) in the comparison group. By applying guidelines set forth by What Works Clearinghouse (2014), it was observed that the overall attrition of 9% and differential attrition of 6% falls into a category of low attrition, which is operationalized as a condition where the balance between overall and differential attrition, “. . . is expected to result in an acceptable level of bias even under the conservative assumptions” (p. 12).

#### Procedures

**Screening and assignment.** Research staff screened all consented fourth grade students at the 16 schools during the fourth or fifth week of school using the reading comprehension subtest of the GMRT. All students scoring at or below the 30th percentile on this measure were identified for the study and randomly assigned within school to treatment (Passport;  $n = 226$ ) or comparison ( $n = 225$ ) using stratification on the screening measure.

Students assigned to the treatment group were subsequently assigned within school to small groups of four to seven students (a total of 43 groups across schools). Each treatment group received the Passport to Literacy intervention daily for 30 min sessions for 25 weeks. Students assigned to the comparison group received the typical services provided by the school.

**Data collection.** Following screening, pretest measures were administered at the end of September and beginning of October to all participants. Posttest assessments were administered in early May, within 2 weeks of the intervention completion. Assessments were counterbalanced by measure and were administered by trained research assistants blind to condition and assignment. Prior to pretesting and posttesting, assessment staff were required to demonstrate 100% accuracy in administration and scoring on all measures. Further, all measures were double-scored and double-entered by two, independent research staff.

We observed students' school provided reading instruction. First, we collected data on students' core, classroom reading instruction (Tier 1) in the fall and in the spring to understand the type and amount of reading instruction students received in their classrooms. Observers were trained to use the Instructional Content Emphasis Instrument-Revised (ICE-R; Edmonds & Briggs, 2003) to record what was taught, how long it was taught, and the instructional grouping used for teaching. Following the guidelines of the ICE-R, specific instructional activities were coded if they lasted for at least 1 min. Content categories included phonemic awareness, phonics/word recognition, fluency, vocabulary/oral language development, comprehension, spelling, text reading sep-



arate from other instruction, and nonliteracy activities (e.g., other academic instruction, noninstructional time). Observers also coded instructional groupings as whole class, small-group, pairs, independent activity/assignment, or individualized instruction. Student engagement for the overall observation was coded using a three-point rubric (3 = *high engagement*, 1 = *low engagement*). Finally, observers assigned a global quality of instruction rating for the overall observation based on a 4-point Likert scale ranging from 1 (*weak*) to 4 (*excellent*). This global instructional quality variable considered a teacher's use of direct and explicit language, modeling, students' opportunities for practice, specific feedback, monitoring and encouragement of engagement, scaffolding of tasks, and pacing throughout the lesson.

We used a multiple-step training process to establish interrater reliability for the Tier 1, classroom reading instruction observations in fall and again in the spring before each round of observations began. Initially, each observer was instructed on the meaning of each code/indicator and provided specific examples. Next the coding process was modeled by the principal investigator of the project using a short video segment of reading instruction from another project. Finally, each observer practiced coding using several novel video segments that were subsequently discussed with the principal investigator. Each observer established 90% or higher coding accuracy with the principal investigator (i.e., gold standard approach) on a separate video segment of reading instruction. Observers reestablished reliability prior to spring observations with new video segments. All coders were required to be above 90% reliability at each time point. Exact interrater reliability across coders and time periods was 95.1%.

To identify any supplemental reading instruction/intervention, research staff completed brief interviews with classroom teachers regarding additional reading support beyond core reading instruction for each participating student. Each semester teachers indicated the session time, frequency, grouping, implementer, and implementer's credentials. All supplemental intervention sessions in both study conditions were audio recorded at three time points during the school year (fall, winter, and spring); recordings of instruction were then coded using the ICE-R measure to describe any interventions students received.

In addition, the fidelity of implementation of the Passport to Literacy intervention was monitored monthly via direct observations of lessons with a measure specific to the required components of the Passport to Literacy intervention. Interventionists were observed and scored on implementation of each activity, student academic engagement, and quality of instruction for each lesson component. The scale for implementation ranged from 0 (*teacher did not complete elements of component*) to 3 (*all or nearly all required elements completed*), while engagement and instructional quality were also rated from 1 (*weak engagement or quality*) to 3 (*excellent engagement or quality*). Instructional quality indicators included ongoing monitoring, redirection of off-task behavior, positive and corrective feedback, organization of materials, and appropriate selection of additional items for practice when needed. Each observer obtained a minimum reliability of 90% in comparison to a gold standard rating by the project coordinator prior to formal data collection; across three observers, reliability was 95.3%.

## Description of Instruction

**Tier 1, classroom reading instruction.** Data from observations of core reading instruction received by all participating students indicated that the length of reading classes was, on average, 75.40 min ( $SD = 26.34$ ). Within this instruction, activities devoted to reading comprehension and vocabulary development were most prevalent, accounting for nearly 35 min (46%) of total time. Instruction devoted to word analysis/decoding was minimal ( $< 1$  min [ $< 1\%$  of time]), while time spent in reading of connected text and/or reading fluency practice was approximately 9 min (12% of time) daily. Of note, approximately 15 min (20% of time) was spent in differentiated instructional activities where students in the class were engaged in different activities simultaneously. The additional 14 min (19%) of time was spent in other types of activities (e.g., transitions). Core reading instruction primarily occurred as whole-class instruction (approximately 45 min or 60% of time on average). Just less than 10 min (13%) of instructional time consisted of students working independently on the same activity, while approximately 8 min (11%) was spent in either small-group or paired instructional activities. Generally, the global ratings of instruction for the core classroom instruction were suggestive of high average instructional quality ( $M = 3.17$ ,  $SD = .59$ ). Similarly, academic engagement by students during core reading instruction was rated as high ( $M = 2.78$ ,  $SD = .55$ ).

**School-provided supplemental instruction.** A total of 130 students ( $n = 62$  treatment [27%];  $n = 68$  comparison [30%]) also received supplemental intervention provided by their respective schools for all or part of the year. Teacher reports indicated that this supplemental reading intervention was most often delivered by classroom teachers (20%) or other certified teachers (43% of students) with eight interventions (18%) delivered by a paraprofessional or a volunteer, and 6 interventions (14%) delivered by speech-language pathologists. Interventions most often held sessions between 31 and 50 min (70%) with 16% of the interventions meeting between 21 and 30 min and 10% between 10 and 20 min. Seventy percent of the interventions were held in group sizes of one to five students. Nine students received two supplemental interventions during the school day.

Across the 2 years, based on recordings of this instruction, intervention sessions averaged 28.34 min ( $SD = 13.78$ ). The most frequent instructional activities involved those related to comprehension of text ( $M = 8.27$  min,  $SD = 7.60$ ) with about 29% of intervention time, as well as vocabulary and oral language development ( $M = 4.45$  min,  $SD = 5.90$ ) for about 16% of intervention time. Text reading without other instruction occurred for approximately 6 min ( $M = 6.43$  min,  $SD = 5.1$ ) or 23% of intervention time, and students received phonics/decoding instruction for an average of 3.84 min ( $SD = 7.86$ ) or 14% of intervention time. Minimal instruction (0–4% of intervention time) was focused on oral reading fluency practice ( $M = .53^{\dagger}$  min,  $SD = 1.71$ ), spelling ( $M = 1.22$  min,  $SD = 3.27$ ), or phonemic awareness ( $M = .04$  min,  $SD = .23$ ). During the additional reading intervention, an average of 1.86 min ( $SD = 3.74$ ) or 7% of instructional time was spent in other academic instruction. About 4% of the intervention time was spent in noninstructional activities ( $M = 1.04$  min,  $SD = 3.68$ ). The mean rating of instructional quality for students who received supplemental reading instruction was 2.83 ( $SD = .47$ ) and student engagement was also high ( $M = 2.65$ ,  $SD = .36$ ).



Table 1 provides information on this typical school instruction in comparison to the treatment intervention sessions.

**Passport to Literacy intervention.** We provided the standard implementation of the Passport to Literacy intervention program at the fourth-grade level to students in the treatment condition. Passport to Literacy is designed to be used as a supplemental reading intervention provided in small groups daily for 30 min sessions for 1 school year (up to 120 lessons). We scheduled the intervention sessions with the school/teachers outside of their core, classroom reading instruction block, typically during the time that schools had already designated for intervention/enrichment.

The Passport to Literacy intervention is broken into 12, 10-day adventures, with each lesson targeting phonics and word recognition, fluency, vocabulary, and comprehension. To monitor students' mastery of content and progress on oral reading fluency, checkpoints are designed at the fifth and 10th lesson of each adventure. The sequence of instruction began with an *Adventure Starter* activity (approximately 3–5 min) to build background knowledge by linking the lessons and readings to the adventure. Then, lessons included two major components; the first, *Word Works*, or word study, taught students to read and understand unknown multisyllabic words using strategies to break words down into smaller parts, including affixes, roots, and syllabication. For the first 6 weeks, the Word Works instruction was 20 min and also included more basic word reading skills such as letter/sound identification, decoding, sight word reading, word families, and spelling instruction. In subsequent lessons, Word Works was reduced to 5 min, but also included a brief 2 min *Warm-Up* where students received additional word study practice through review and application of previously learned letter combinations, sight words, spelling rules, and word endings.

Then, during the second component, *Read to Understand*, students were taught the meaning of vocabulary words introduced during Word Works, as well as comprehension skills and strategies to apply while reading fiction and nonfiction. For example, lessons offered explicit instruction in previewing, setting purpose, text structure and evaluation, making inferences and taking perspectives, drawing conclusions, author's purpose, sequencing, main idea, summarizing, independent reading fix-up strategies, teacher and reader questioning, and making connections within and across texts. In the first 6 weeks, instruction in the Read to Understand component lasted 10 min and in subsequent lessons, was increased

to 25 min. Lessons also included a brief focus on fluency (reading with appropriate accuracy, rate, and expression) during the text reading.

**Intervention teachers and training.** A total of 17 teachers, hired by the research team, were responsible for teaching the Passport to Literacy lessons. All the teachers had a bachelor's degree, four (33.3%) had obtained a master's degree in education, and one had a PhD. Twelve of the interventionists were certified teachers and one was a counselor. The other four had degrees in noneducation areas. All intervention teachers were female. Three teachers identified themselves as Hispanic ethnicity. In terms of race, 11 (65.7%) teachers were White and five teachers (29.4%) were Black and one chose not to fill in the information.

Prior to the start of instruction, intervention teachers participated in approximately 8 hr of training over the course of 2 days. Training provided by the project coordinators at each site, allowed interventionists to become oriented to the project, familiarize themselves with the Passport to Literacy intervention program and instructional routine, practice implementation of lessons, and discuss positive behavior supports. Once intervention sessions with students were initiated, twice monthly coaching visits were conducted by the project coordinators. These visits allowed teachers to receive feedback on implementation as well as discuss any questions or concerns. Finally, monthly meetings with all intervention teachers were held at each site to provide continued support and ensure fidelity of implementation.

**Intervention implementation and fidelity.** The total number of Passport to Literacy lessons covered for each of the intervention groups ranged from 83 to 106 sessions. For those individual students who remained in the school for the duration of the intervention, the number of lessons attended ranged from a low of 58 sessions to a high of 106 sessions ( $M = 93.79$ ,  $SD = 7.82$ ).

As noted earlier, each intervention teacher recorded three intervention lessons during the year, and these recordings were coded for instructional content and quality using the ICE-R to directly compare the instructional elements in Passport and the school-provided interventions. On average, the treatment session instruction was 28.56 min ( $SD = 4.07$ ) in length. Instruction focused on developing students' reading comprehension ( $M = 11.80$  [41% of intervention time],  $SD = 5.65$ ) and vocabulary/oral language ability ( $M = 6.05$  [21% of intervention time],  $SD = 4.81$ ). During treatment lessons, students engaged in text reading for 4.72 min ( $SD = 2.43$ ) or 17% of intervention time, decoding and word reading activities for 3.29 min ( $SD = 3.11$ ) or 12% of intervention time and practiced spelling for just over 1 min ( $M = 1.32$ ,  $SD = 2.34$ ) or 5% of intervention time. Explicit instruction in oral reading fluency was observed for 0.26 min ( $SD = 0.92$ ) or 1% of intervention time, on average. During treatment lessons, less than 1 min (1%) of time was considered either noninstructional in nature ( $M = 0.18$ ,  $SD = 0.64$ ) or focused on instruction in another academic area such as writing or grammar ( $M = .27$ ,  $SD = 0.83$ ). Ratings of instructional quality indicated high-average quality ( $M = 3.37$ ,  $SD = .62$ ) and on average, intervention students were engaged during instruction ( $M = 2.85$ ,  $SD = .43$ ).

In terms of direct fidelity of implementation to the Passport to Literacy lessons, mean implementation ratings for each tutor implementation were high, ranging from 2.71 to 3.00, across the lesson components. Similarly, mean ratings of student academic

Table 1  
Average Intervention Instructional Time in Minutes and  
Percentage of Time by Study Condition

Instructional component	Passport intervention		School-provided intervention	
	No. of min	% of total time	No. of min	% of total time
Phonics and word recognition	3.29	12	3.84	14
Spelling	1.32	5	1.22	4
Reading fluency	.26	1	.53	2
Vocabulary/oral language	6.05	21	4.45	16
Comprehension	11.80	41	8.27	29
Non-instructional text reading	4.72	17	6.43	23
Other academic instruction	.27	1	1.86	7
Noninstruction	.18	1	1.04	4

engagement (2.85 to 3.00) and quality of tutor instruction (2.76 to 3.00) for each component were high.

Dependent Measures

Project staff blind to condition assessed students’ word reading, decoding, vocabulary, reading fluency, and reading comprehension in the fall and spring. Because of the high correlation between students’ word reading and oral reading fluency (see Table 2), we included only the word reading measures in the dependent variables, but examined possible moderation of students’ fluency on outcomes.

**Woodcock-Johnson III Tests of Achievement (WJIII; Woodcock, McGrew, & Mather, 2001).** To assess word reading and comprehension, we selected four individually administered subtests from the nationally standardized WJIII. The letter–word identification subtest measures recognition of real words, and begins with individual letters. The word attack subtest measures decoding skill and includes items that are pseudowords, which begin with a few single letter sounds and progress to decoding of complex pseudowords. The picture vocabulary test asks students to name pictured objects increasing in difficulty. The passage comprehension subtest measures how well students can read text with missing words, presented as a cloze procedure in which students read the sentences silently and are asked to supply the missing word. Test authors report that test–retest reliability for these four subtests at fourth grade are .81, .85, .77, and .86, respectively.

**Dynamic Indicators of Basic Early Literacy Skills–6th Edition (DIBELS; Good & Kaminski, 2002).** To assess student’s ability to read connected text with speed and accuracy, we administered the oral reading fluency (ORF) subtest from DIBELS. Students read three separate passages aloud for 1 min and the total number of correct words read per minute from the passage is considered the oral reading fluency rate. Test-retest reliabilities for ORF with elementary age students range from .92 to .97; alternate-form reliability across passages from the same level is reported as .89 to .94 (Good et al., 2004).

**GMRT (MacGinitie et al., 2006).** The GMRT is a group-administered, norm-referenced test. We administered the vocabulary and comprehension subtests. The fall reading comprehension scores were used to screen students for inclusion in the study. Vocabulary presents words in context. The student chooses the correct meaning of the target word. Comprehension provides students with reading passages and multiple choice questions. Questions address facts, inferencing, and drawing conclusions. Test–retest reliabilities are above .85. Construct validity estimates range from .79 to .81.

Analytic Approach

For both research questions, a longitudinal, multilevel structural equation modeling (ML-SEM) framework was used to estimate primary and conditional impacts. A structural equation model approach is useful as it minimizes the limitation of measurement error inherent to individual observed measures by leveraging the common variance across multiple assessments of a construct. Common specifications of the ML-SEM for randomized controlled trials include latent factors of pretest and posttest measures at both a lower level unit, such as students, and at an upper level unit (e.g., classrooms). Similar to multilevel models of observed outcomes, the ML-SEM includes the regression of posttest on pretest but in this case with latent variables. Estimation of the treatment effect may occur through one of two common approaches. One methodology includes the simple regression of the posttest on *k*-1 dummy codes for a grouping variable, where *k* is the number of treatment arms, to reflect whether an individual received the intervention or not. An alternative approach does not include a variable for treatment status, but rather tests for group differences through a multiple group invariance approach. In this instance the test of impact is estimated by inspecting the posttest means for invariance between groups when constraining other parameters of the model to be equal (e.g., loadings, residual variances, regression of posttest on pretest). The difference in standardized posttest means between

Table 2  
Descriptive Statistics and Correlations for Study Measures

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Fall GMRT RC	—												
2. Fall WJ PC	.32	—											
3. Fall WJ LWID	.30	.60	—										
4. Fall WJ WA	.26	.52	.77	—									
5. Fall GMRT Voc	.39	.49	.52	.41	—								
6. Fall WJ PV	.14	.51	.25	.12	.33	—							
7. DIBELS ORF	.29	.51	.70	.62	.46	.13	—						
8. Spring GMRT RC	.32	.46	.38	.32	.43	.23	.44	—					
9. Spring WJ PC	.35	.64	.54	.43	.50	.43	.47	.47	—				
10. Spring WJ LWID	.29	.60	.82	.72	.49	.21	.69	.39	.61	—			
11. Spring WJ WA	.24	.49	.76	.76	.44	.19	.60	.30	.50	.79 <sup>‡</sup>	—		
12. Spring GMRT Voc	.31	.55	.51	.41	.64	.34	.49	.64	.53	.54	.46	—	
13. Spring WJ PV	.17	.52	.33	.16	.39	.74	.23	.26	.54	.36	.26	.43	—
<i>M</i>	440.61	481.92	484.78	490.32	445.93	486.44	80.35	456.69	487.54	493.01	495.90	462.06	491.11
<i>SD</i>	19.37	12.16	18.97	16.55	27.51	12.41	26.87	24.13	9.66	17.85	14.40	30.67	11.91
<i>N</i>	412	409	409	409	328	409	410	405	404	404	404	406	404
% missing data	.0%	.7%	.7%	.7%	20.4%	.7%	.5%	1.9%	1.9%	1.9%	1.9%	1.5%	1.9%

Note. GMRT RC = Gates-McGinitie Reading Comprehension; WJ PC = WJ-III Passage Comprehension; WJ LWID = WJ-III Letter Word Identification; WJ WA = WJ-III Word Attack; GMRT Voc = Gates-McGinitie Vocabulary; WJ PV = WJ-III Picture Vocabulary. All correlations statistically significant at least *p* < .05.



groups then represents the standardized effect size difference. ML-SEMs have received fair attention in the literature as of late (e.g., Goddard, Goddard, Kim, & Miller, 2015; Heck & Thomas, 2015) as a method to not only overcome measurement issues but also in increasing power to detect effects due to latent variables increasing reliability of the measured construct. A known property of effect sizes is that they are negatively related to unreliability of measurement. Subsequently, with greater precision in measurement through the latent variable, it is possible to detect larger effects that may not be possible with observed variable error.

Despite the increasing prevalence of ML-SEM in the literature for testing treatment effects, a limitation in application has been to randomized designs where not all units are nested. In partially nested randomized controlled trials (PN-RCT; Baldwin, Bauer, Stice, & Rohde, 2011; Lohr, Schochet, & Sanders, 2014), only some individuals are nested within a group. For the present study, the partial nesting is observed where students receiving the intervention were all nested within small groups but the comparison students were not. Baldwin et al. (2011) noted that in their review of studies with PN-RCT designs, researchers frequently ignored this structure to the detriment of standard error estimation. Although robust methods have been proposed that model observed measures for PN-RCT designs, less attention has been given to the

treatment of PN-RCT data in the ML-SEM context. Sterba et al. (2014) presented an approach within Mplus that allows an individual to match the ML-SEM methodology to the PN-RCT design. However, a limitation of reported approaches for observed and latent variable approaches for PN-RCT data is that they involve the introduction of ancillary variables into the data, as well as additional model specifications (e.g., adjusting estimation of the denominator degrees of freedom for observed variables) that are not possible to implement across commonly used software.

A more naturalistic approach to treating PN-RCT data is to view the nesting structure through  $n$ -level SEM ( $n$ SEM; Mehta & Neale, 2005) which easily accommodates complex nesting. Within  $n$ SEM, observed and latent variables may be used across multiple levels. The concept of *level* in  $n$ SEM takes on unique meaning differing from multilevel modeling. That is, a level typically refers to a unit of clustering for one set of observations within another unit such as students nested within classrooms. A level in  $n$ SEM refers to this type of nesting but further describes any meaningful, nominal grouping of individuals such as male or female, students eligible for free/reduced lunch or not, or those who received an intervention or not. This more flexible use of level allows us to more naturalistically situate the PN-RCT design in the  $n$ SEM framework. Consider a sample  $n$ SEM model in Figure 1 that is

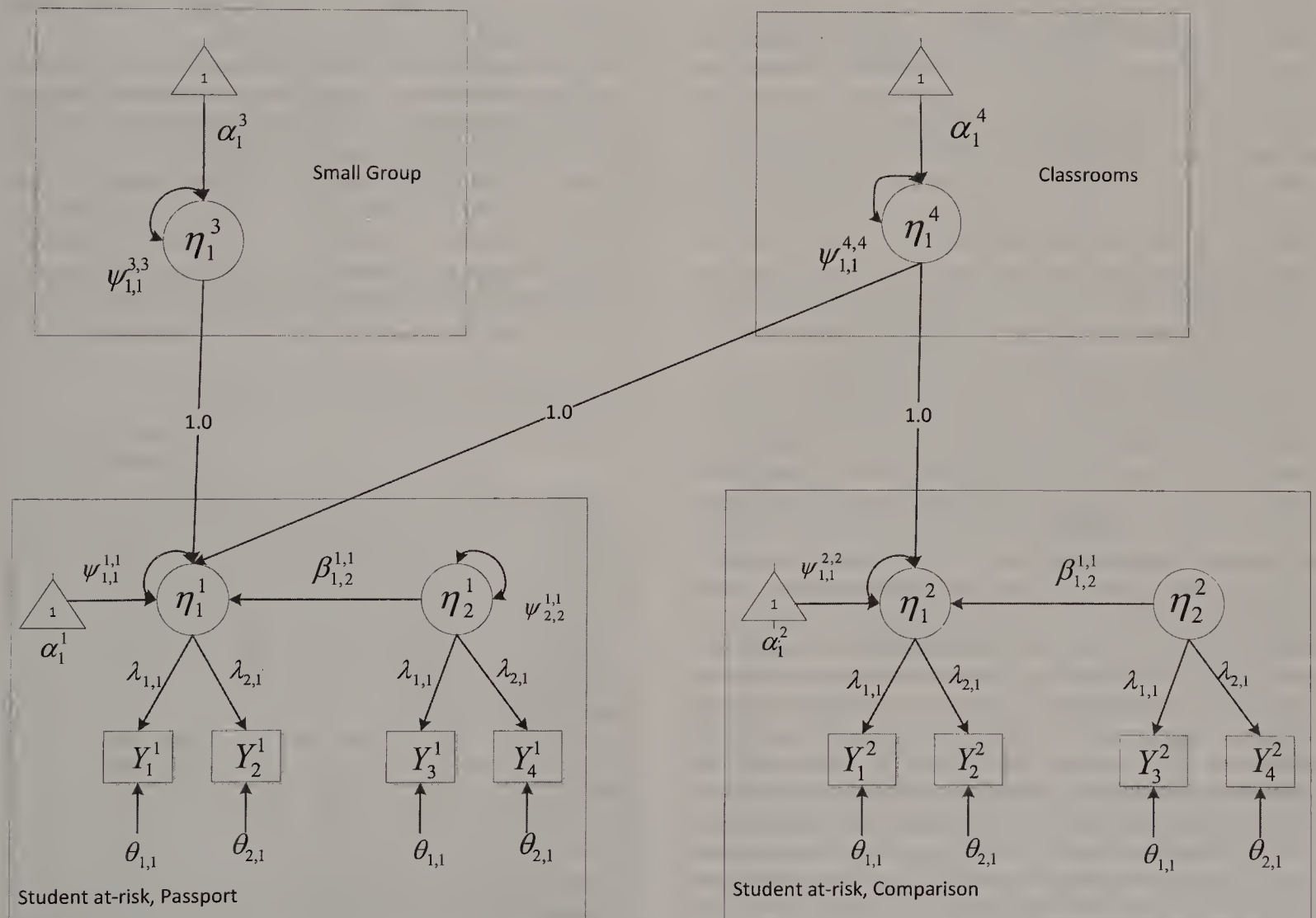


Figure 1. Sample  $n$ -level structural equation measurement model for partially nested designs.

relevant to the current study. Note that there are four boxes that are each representative of participant groupings. Pertaining to students, there are two levels of groupings one for the Passport students (Level 1) and one for comparison students (Level 2). Small group represents a nesting structure for only the Passport students (Level 3) and Classrooms represent the nesting of students from both student groups in classrooms (Level 4). Figure 1 then represents a four-level partially nested, cross-classified SEM where the comparison students are nested within classrooms and the Passport students are cross-classified by small groups and classrooms.

At this point, it may useful to provide an introduction to more specific components of the model. For both the Passport and comparison levels, the SEM specifies that there is a posttest ( $\eta_1^1$  for Passport and  $\eta_1^2$  for comparison), where the superscript notation denotes the level for the parameter and the subscript denotes the parameter number. Thus,  $\eta_1^1$  is the first Level-1 latent variable, (i.e., the Passport posttest latent variable) and  $\eta_1^2$  is the first Level-2 latent variable for the comparison group at the posttest.  $\eta_2^1$  then is the second latent variable for the Passport group (i.e., the pretest) and  $\eta_2^2$  is the pretest latent variable for the comparison group. The latent variables in Passport are indicated by the four measures  $Y_1^1$  to  $Y_4^1$ , two at pretest and the same two at posttest, as are the latent variables for comparison group indicated by the same measures  $Y_1^2$  to  $Y_4^2$ . Each of the observed measures has a residual ( $\theta$ ) and loading ( $\lambda$ ). Note that the loading subscripts are the same from posttest to pretest and between the Passport and comparison groups. This specification denotes that the model constrains the estimated values to be equal across groups, as it does also for the residual variances and the regression of the posttest latent construct on the pretest ( $\beta$ ). Across all four levels, there are latent means ( $\alpha$ ) and variances ( $\psi$ ). As a multilevel model, only the latent means at the student levels (i.e., Passport and comparison) are estimated; they are fixed at 0 at the small group and classroom levels. Similar to a longitudinal SEM, the pretest means (not reflected in the diagram) are set at 0 and the variances are fixed at 1. This specification is so that the means at the posttest are standardized such that the difference between  $\alpha_1^1$  and  $\alpha_1^2$  is the standardized treatment effect.

The model building process for the PN-RCT *n*SEM occurred in two phases with four models each. Phase 1 was focused on testing longitudinal invariance of the loadings and intercepts, and Phase 2 tested between-level posttest invariances. Within Phase 1, three models were tested: (1) freed loadings and intercepts across pretest and posttest latent variables in treatment and comparison groups (Model 1); (2) invariant loadings and freed intercepts across pretest and posttest latent variables in treatment and comparison groups (Model 2); (3) invariant loadings and intercepts across pretest and posttest latent variables in treatment and comparison groups (Model 3). These steps were necessary to evaluate whether a fully invariant model for intercepts and loadings was plausible such that the latent means are reflective of actual latent mean differences and not loading/intercept structure differences. For Phase 2, five models were tested to test for posttest invariance across combinations of the treatment, comparison, and small group levels: (1) freed loadings and intercepts across treatment, comparison, and small group levels (Model 4); (2) invariant loadings and freed intercepts between treatment and comparison levels (Model 5); (3) invariant loadings and intercepts between treatment and comparison levels (Model 6); (4) invariant loadings and freed

variances between treatment and small group levels (Model 7); and (5) invariant loadings, intercepts, pretest means, and variances across treatment, comparison, and small group levels (Model 8). Each set of eight models were applied to reading comprehension, word reading, and vocabulary outcomes. Exploratory analyses in the study tested whether EL status, pretest, letter-word identification, or oral reading fluency moderated the relation between treatment status and posttest performance. Model comparisons were made using the deviance statistic as well as the Aikake information criterion and Bayesian information criterion indices. A log-likelihood difference test was used for hypothesis testing of model differences.

Results

Descriptive Statistics and Correlations

A preliminary review of the data for missingness (see Table 2) showed that complete data were available for the fall GMRT-RC measure ( $n = 412$ ), but missing data rates varied from .7% to 20.4% for other measures. The reason for the high level of missing data on the fall GMRT vocabulary measure was it was not administered in one site in Year 1. Little's missing completely at random (MCAR) test suggested that all missing data met reasonable assumptions for MCAR,  $\chi^2(81) = 77.99, p > .500$ ; thus, using full information maximum likelihood for model estimation was appropriate and would not negatively bias results.

Table 2 presents the full sample student performance results on the individual measures of reading comprehension, word reading, and vocabulary at fall and spring and Table 3 reports means and standard deviations by treatment condition. Students' scores on the measures were consistently higher at the spring compared to fall. Correlations among the measures in the fall ranged from .12 between WJIII picture vocabulary and word attack to .77 between WJII word attack and letter-word identification. Spring correlations ranged from .26 between WJIII picture vocabulary and

Table 3  
Descriptive Statistics of Measures by Condition

Measure	Passport			Comparison		
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>
Fall GMRT RC	199	439.96	19.96	213	441.23	18.82
Fall WJ PC	199	481.52	11.67	210	482.30	12.61
Fall WJ LWID	199	484.43	18.82	210	485.12	19.14
Fall WJ WA	199	488.91	16.99	210	491.65	16.03
Fall GMRT Voc	159	444.87	27.09	169	446.93	27.93
Fall WJ PV	199	486.85	12.98	210	486.05	11.84
Fall DIBELS ORF	198	78.11	25.58	212	82.44	27.91
Spring GMRT RC	198	459.25	23.93	207	454.23	24.11
Spring WJ PC	198	488.12	9.35	206	486.98	9.93
Spring WJ LWID	198	492.79	17.14	206	493.23	18.54
Spring WJ WA	198	495.47	14.67	206	496.31	14.21
Spring GMRT Voc	198	462.08	31.87	208	462.04	29.55
Spring WJ PV	198	491.70	11.97	206	490.54	11.85

Note. GMRT RC = Gates-MacGinitie Reading Comprehension; WJ PC = WJ-III Passage Comprehension; WJ LWID = WJ-III Letter Word Identification; WJ WA = WJ-III Word Attack; GMRT Voc = Gates-MacGinitie Vocabulary; WJ PV = WJ-III Picture Vocabulary; ORF = Oral Reading Fluency.



Table 4  
*Confirmatory Factor Analysis Model Fit Comparison for Latent Reading Comprehension, Word Reading, and Vocabulary*

Outcome	Model	−2LL	df	AIC	BIC	Δ-2LL	Δdf	p
Reading comprehension	1	6766.20	12	6790	6847	.65	2	.723 <sup>a</sup>
	2	6766.21	11	6788	6840			
	3	6766.85	9	6784	6827			
	4	6578.35	18	6614	6698			
	5	6578.35	17	6612	6692	4.18	6	.652 <sup>b</sup>
	6	6578.36	16	6610	6686			
	7	6578.39	18	6612	6692			
	8	6582.57	12	6606	6662			
Word reading	1	6650.71	12	6675	6731	.3	2	.861 <sup>a</sup>
	2	6650.22	11	6672	6724			
	3	6650.56	9	6673	6716			
	4	6364.37	18	6400	6485			
	5	6364.36	17	6398	6478	1.72	5	.886 <sup>b</sup>
	6	6364.62	16	6397	6472			
	7	6364.37	17	6398	6478			
	8	6366.09	12	6390	6446			
Vocabulary	1	6283.66	12	6308	6363	.87	2	.647 <sup>a</sup>
	2	6283.65	11	6305	6356			
	3	6284.52	9	6303	6344			
	4	6933.29	18	6969	7054			
	5	6933.29	17	6967	7047	.78	5	.978 <sup>b</sup>
	6	6933.47	16	6965	7041			
	7	6933.29	17	6967	7047			
	8	6934.07	12	6958	7014			

*Note.* −2LL = −2\*log likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion. Model 1 = treatment-comparison, pretest-posttest freed loadings and intercepts; Model 2 = treatment-comparison, pretest-posttest, invariant loadings, freed intercepts; Model 3 = treatment-comparison, pretest-posttest, invariant loadings and intercepts; Model 4 = treatment-comparison -small group freed loadings and intercepts; Model 5 = treatment-comparison invariant loadings, freed intercepts; Model 6 = treatment-comparison invariant loadings and intercepts; Model 7 = treatment-small group invariant loadings, freed variances; Model 8 = treatment-small group-comparison invariance loadings, intercepts, means, and variances.  
<sup>a</sup> Model is compared with Model 2. <sup>b</sup> Model is compared with Model 7.

GMRT reading comprehension to .79 between WJII word attack and letter-word identification. Stability coefficients from fall to spring ranged from .32 for GMRT reading comprehension to .82 for WJII letter-word identification, suggesting moderate to high stability in relative rank orders of individuals over time.

Tests of Invariance

Results from the tests of invariance are presented in Table 4. For the first phase of invariance testing, which was related to longitudinal invariance between pretest and posttest between the treatment and comparison groups, results consistently demonstrate that imposing incremental equality constraints on the intercepts and loadings did not significantly denigrate fit. This step is important as it suggests that the means and loadings didn't differ by forcing them to be equal across groups. For reading comprehension, the difference in deviance between Models 2 and 3 was negligible (Δ−2LL = 0.65) and not statistically significant (*p* = .723). Similarly, no significant differences were observed between Models 2 and 3 for word reading (Δ−2LL = 0.30, *p* = .861) or vocabulary (Δ−2LL = 0.87, *p* = .647). Phase 2 invariance testing in the posttest invariance among the treatment, comparison, and small groups (Models 4 through 8) show that no substantive difference was observed in the deviance statistic. In fact, the largest difference in deviance between Model 4 (the least restric-

tive model) and Model 8 (the most restrictive model) was for reading comprehension where the deviance difference was <4 points with six degrees of freedom, a nonsignificant finding. When comparing the final two models, no significant differences were observed for reading comprehension (Δ−2LL = 4.18, *p* = .652), word reading (Δ−2LL = 1.72, *p* = .886), or vocabulary (Δ−2LL = 0.78, *p* = .978).

nSEM Primary Impact Model Results

Primary impact model results for the three latent outcomes of reading comprehension, word reading, and vocabulary related to the first research question are presented in Figures 2 and 3. Using a similar methodology for comparing the factor analytic models, the impact analyses tested constrained and freed estimate versions of the nSEM in Figure 1. In the constrained version of the model, the latent posttest means for the Passport and comparison groups (i.e., α<sub>1</sub><sup>2</sup> and α<sub>2</sub><sup>2</sup>; Figure 1) were constrained to be equal. This constraint was relaxed for a second model test of mean difference. A log-likelihood difference test was used for hypothesis testing of model differences. The model comparison for reading comprehension (see Table 4) showed that the model with freed posttest means fit better than the model with constrained means (Δ−2LL = 9.47, Δ*df* = 1, *p* < .001). Figure 2 shows that controlling for the pretest relation to posttest (β = 1.08), the standardized mean posttest

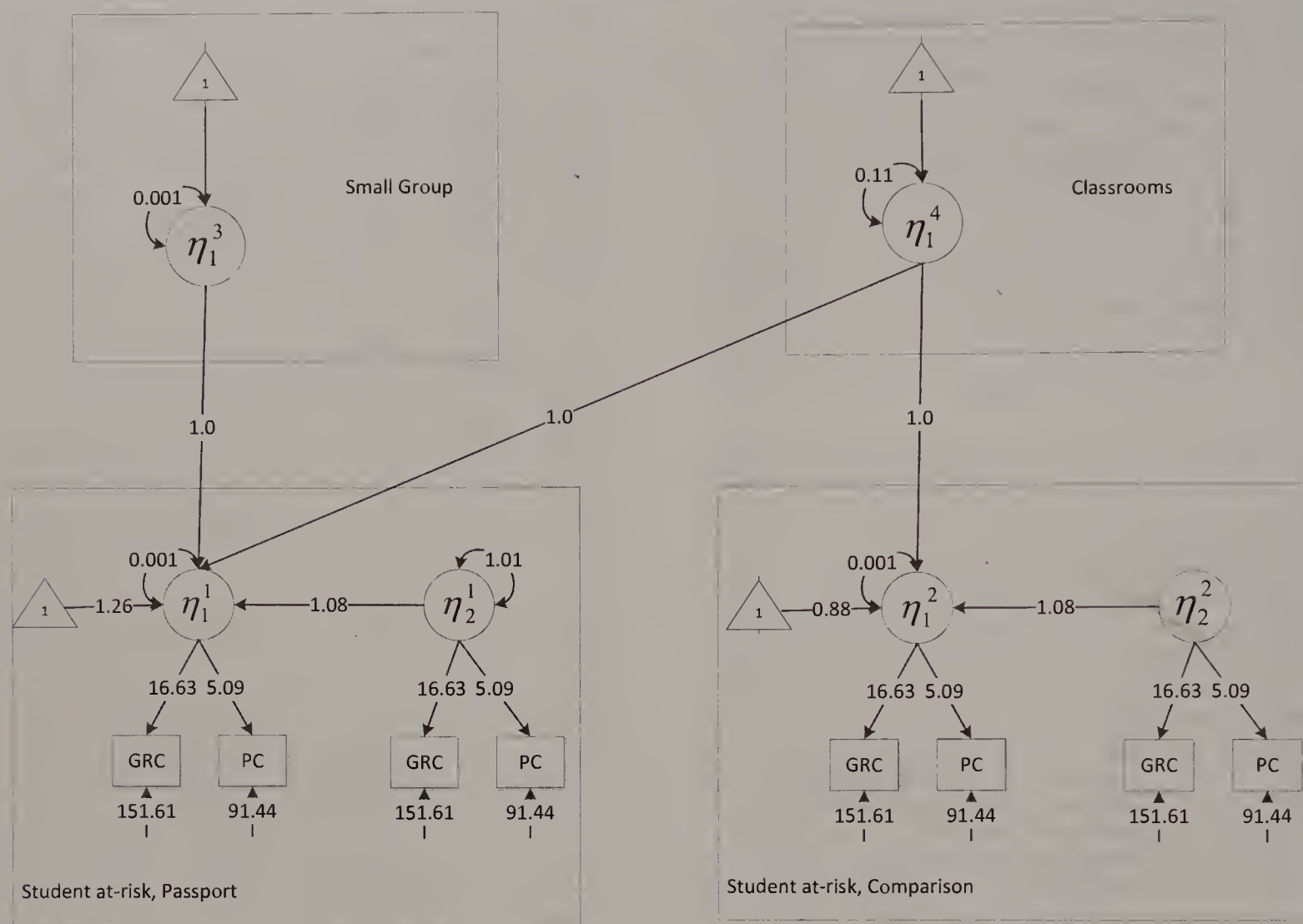


Figure 2. Primary impact  $n$ -level structural equation models for partial nested randomized controlled trial for reading comprehension.

value was  $\alpha = 1.26$  for the Passport group and  $\alpha = .88$  for the comparison group, a statistically significant difference. The effect size of Passport for latent reading comprehension outcomes is calculated as the difference between these two scores, or 0.38. No significant differences were observed between the constrained and freed posttest means models for latent word reading ( $p = .280$ ) or latent vocabulary ( $p = .480$ ). Further, no substantive primary impacts for Passport were observed for word reading ( $\Delta\alpha = .06$ ; Figure 3 top), nor was there an impact on vocabulary ( $\Delta\alpha = .08$ ; Figure 3 bottom).

### *n*SEM Exploratory Modeling Results

To address the second research question, exploratory analyses evaluated the moderation of treatment effects based on EL status and selected baseline measures (i.e., pretest, letter-word identification, and oral reading fluency). As previously noted, two methods are frequently employed to test for treatment effects in SEM studies including the inclusion of  $k-1$  dummy codes or multiple groups. In a similar manner, moderation of treatment effects can be tested by including interaction terms in a regression model, or by using the multiple group method. The moderators for our exploratory analyses were a combination of continuous (i.e., baseline/

pretest, letter-word identification, and oral reading fluency) and categorical (i.e., EL). As such, two different approaches were used for tests of moderation.

Three baseline moderation models were tested. The first moderation model, which we call baseline moderation model, tested the impact of the autoregressive, latent pretest construct and whether the relation between latent pretest and posttest varied by group. By releasing the Beta in Figure 1 to be freely estimated for the Passport and comparison groups, and comparing the fit of this model to the primary impact model where the Beta in Figure 1 is constrained to be the same between the two groups, a test is provided as to whether baseline performance moderates the treatment effect. The second and third moderation models, which each used single-item indicators of letter word identification and ORF, was done by first creating a single-item indicator latent construct for the moderator of interest (i.e., where the loading was fixed at 1.0 and the residual variance was set at a reliability adjusted estimate of the sample variance). This factor was set as a predictor of the latent posttest, identical to the Beta parameter in Figure 1, as well as set to covary with the latent pretest for both Passport and comparison groups. Estimation for this type of model required two steps; first, the path from the baseline measure was constrained to



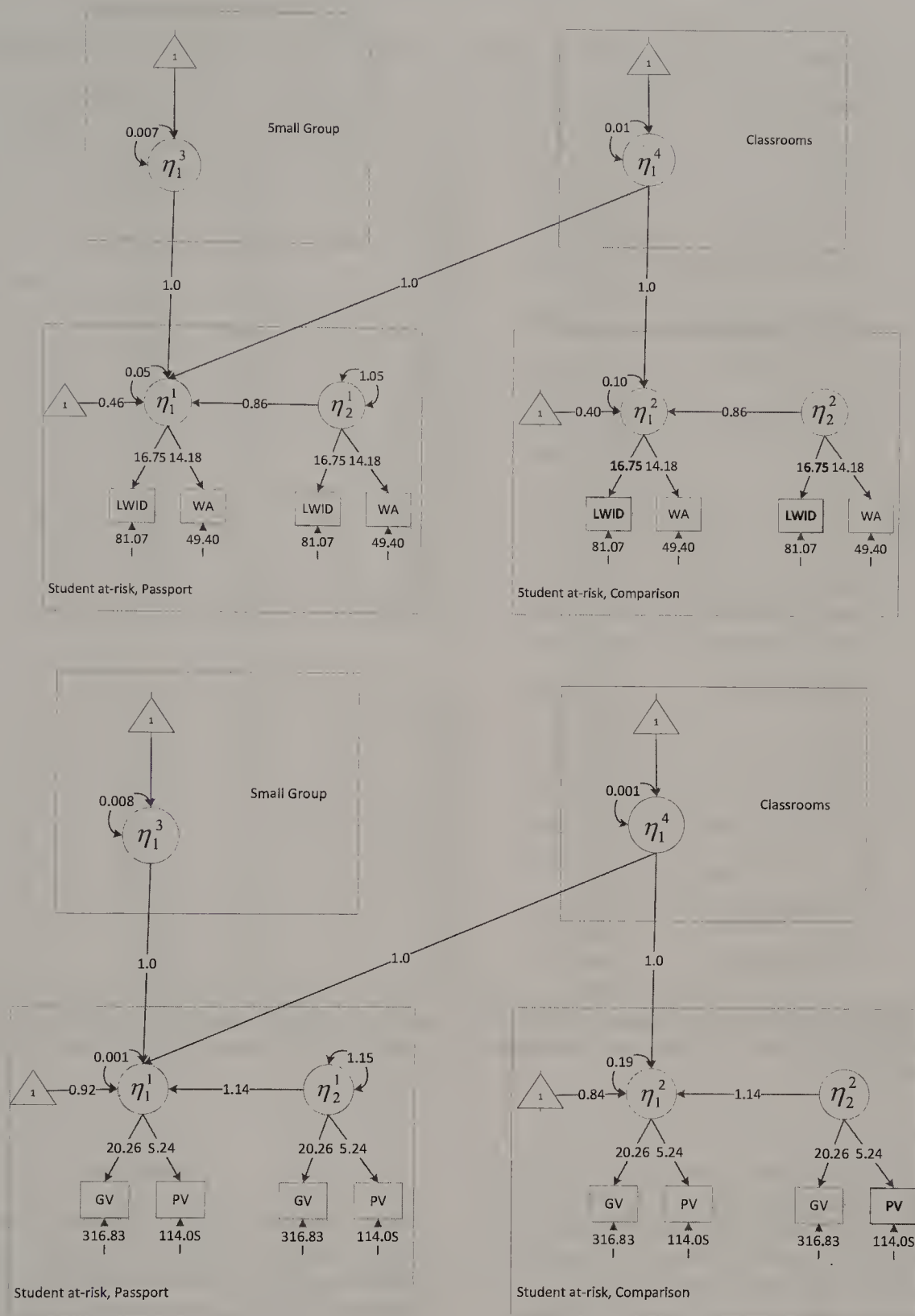


Figure 3. Primary impact  $n$ -level structural equation models for partial nested randomized controlled trial for word reading (top) and vocabulary (bottom).

be equal between Passport and comparison groups. Fit from this model was compared with a model where the Beta constraint was freed for estimation. Improved fit for a freed model provided evidence for moderation.

Results for the three tests of moderation for each outcome are reported in Table 5. For latent reading comprehension, no moderation was observed for baseline latent reading comprehension ( $\Delta-2LL = 0.00, p = 1.00$ ) or baseline oral reading fluency ( $\Delta-2LL = 1.00, p =$

.321), but statistically significant moderation was estimated for baseline letter-word identification ( $\Delta-2LL = 14.87, p < .001$ ) such that students with higher initial word reading scores performed better on reading comprehension in the treatment. No significant moderation was observed for any of the selected moderators for either latent word reading or vocabulary outcomes (see Table 5).

For the EL indicator, moderation was tested by fitting the factor models from Figure 1 separately for EL and non-EL

Table 5

*Fit Comparison for Primary Impact Models and Moderation with EL, Baseline, Letter-Word Identification, and Oral Reading Fluency*

Outcome	Type	Model	-2LL	df	AIC	BIC	$\Delta$ -2LL	$\Delta$ df	p
Reading comprehension	Impact	Constrained	13167.32	16	13199	13285			
		Freed	13157.85	17	13192	13284	9.47	1	.002
	EL moderation	Constrained	3466.47	16	3498	3563			
		Freed	3463.15	17	3497	3566	3.32	1	.068
	Non-EL moderation	Constrained	9654.43	16	9686	9768			
		Freed	9647.69	17	9682	9768	6.74	1	.009
	Baseline moderation	Constrained	13164.84	16	13197	13283			
		Freed	13164.83	17	13198	13290	.01	1	.920
	LWID moderation	Constrained	16545.13	24	16593	16728			
		Freed	16560.00	25	18610	18750	14.87	1	.000
	ORF moderation	Constrained	16875.00	24	16923	17058			
		Freed	16874.00	25	16923	17064	1.00	1	.320
Word reading	Impact	Constrained	12485.20	16	12517	12603			
		Freed	12484.05	17	12518	12609	1.15	1	.284
	EL moderation	Constrained	3323.66	16	3356	3421			
		Freed	3321.07	17	3355	3424	2.59	1	.108
	Non-EL moderation	Constrained	9124.90	16	9157	9239			
		Freed	9124.78	17	9159	9245	.12	1	.729
	Baseline moderation	Constrained	12486.67	16	12519	12605			
		Freed	12486.65	17	12521	12612	.02	1	.888
	LWID moderation	Constrained							
		Freed							
	ORF moderation	Constrained	16032.00	24	16080	16215			
		Freed	16031.00	25	16081	16222	1.00	1	.320
Vocabulary	Impact	Constrained	12826.17	16	12858	12943			
		Freed	12825.67	17	12859	12950	.50	1	.480
	EL moderation	Constrained	3025.1	16	3057	3119			
		Freed	3025.06	17	3059	3125	.04	1	.841
	Non-EL moderation	Constrained	9679.59	16	9712	9793			
		Freed	9678.76	17	9713	9799	.83	1	.362
	Baseline moderation	Constrained	12825.66	16	12858	12943			
		Freed	12825.15	17	12859	12950	.51	1	.480
	LWID moderation	Constrained	16237	24	16285	16418			
		Freed	16235	25	16285	16424	2.00	1	.157
	ORF moderation	Constrained	16550	24	16598	16732			
		Freed	16549	25	16600	16739	1.00	1	.320

*Note.* -2LL =  $-2 \times \log$  likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion; EL = English learner; LWID = letter word identification; ORF = oral reading fluency. LWID moderation was not tested for the latent word reading outcome as it was part of the latent variable itself and included in the pretest construct.

students and evaluating Passport and comparison group posttest mean differences using constrained and freed posttest means similar to the primary impact model. Relevant results for the EL student model (Table 5 and Figure 4) showed no statistically significant difference in posttest means were observed for reading comprehension ( $p = .068$ ), word reading ( $p = .108$ ), or vocabulary ( $p = .841$ ); however, the mean effect size difference in Figure 4 shows small effects in favor of Passport for latent word reading ( $\Delta\alpha = .54 - 0.35 = 0.19$ ) and latent reading comprehension ( $\Delta\alpha = 1.42 - 1.04 = 0.38$ ). No effect of Passport was observed for EL students on latent vocabulary ( $\Delta\alpha = .01$ ). A statistically significant effect of Passport was estimated for non-EL students on reading comprehension ( $p = .009$ ; Table 4) with an effect size of  $\Delta\alpha = .39$  (see Figure 5). No significant effects were estimated for latent word reading ( $p = .729$ ) or vocabulary ( $p = .362$ ); however, different from the other analyses, a small effect on vocabulary was estimated ( $\Delta\alpha = .13$ ; Figure 5).

## Discussion

In this study, our aim was to contribute to the relatively limited body of research on effective comprehensive reading interventions to improve reading comprehension for upper elementary students by extending our prior work examining the effects of a widely used, multicomponent, upper elementary reading intervention. The present study adds uniquely to the existing literature by employing a large sample, using latent variables based on standardized reading measures, and by using a relatively more sophisticated data analytic method ( $n$ SEM) to address differences in nesting within the treatment and comparison groups. In addition, the larger sample also allowed us to examine additional moderators such as initial baseline reading performance and EL status to learn more about for whom the intervention was most effective. The treatment was implemented with a high degree of fidelity that included approximately 94 sessions. Thus, the study is not only rigorous in design, but also is one of the most extensive to date for this grade



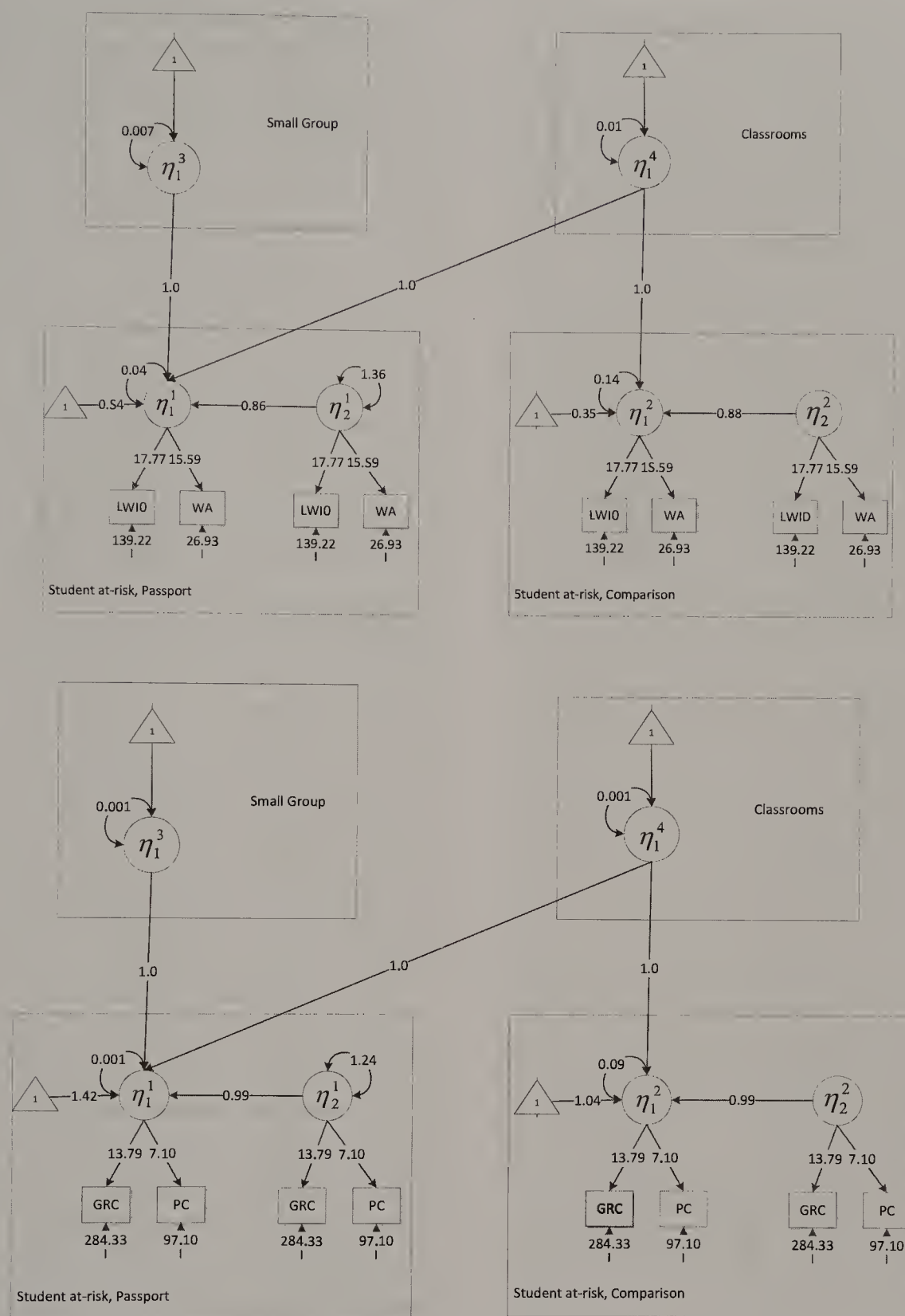


Figure 4. Exploratory  $n$ -level structural equation modeling (SEM) for English Learners on word reading (top) and reading comprehension (bottom)

level; providing a fairly optimal test of the possible effects of implementing this multicomponent intervention at the fourth grade level.

Our first research question addressed main effects of the multicomponent intervention on reading comprehension, word reading, and vocabulary. Consistent with our hypothesis that students with reading difficulties receiving the intervention would outper-

form students receiving only typical school services in reading comprehension, we did find a significant effect of the intervention on reading comprehension with an effect size of 0.38. However, we found no significant effects on word reading ( $ES = 0.05$ ) or on vocabulary ( $ES = 0.08$ ). The magnitude of the effects on comprehension are slightly larger than in our previous study of the Passport to Literacy intervention, which found effect sizes on

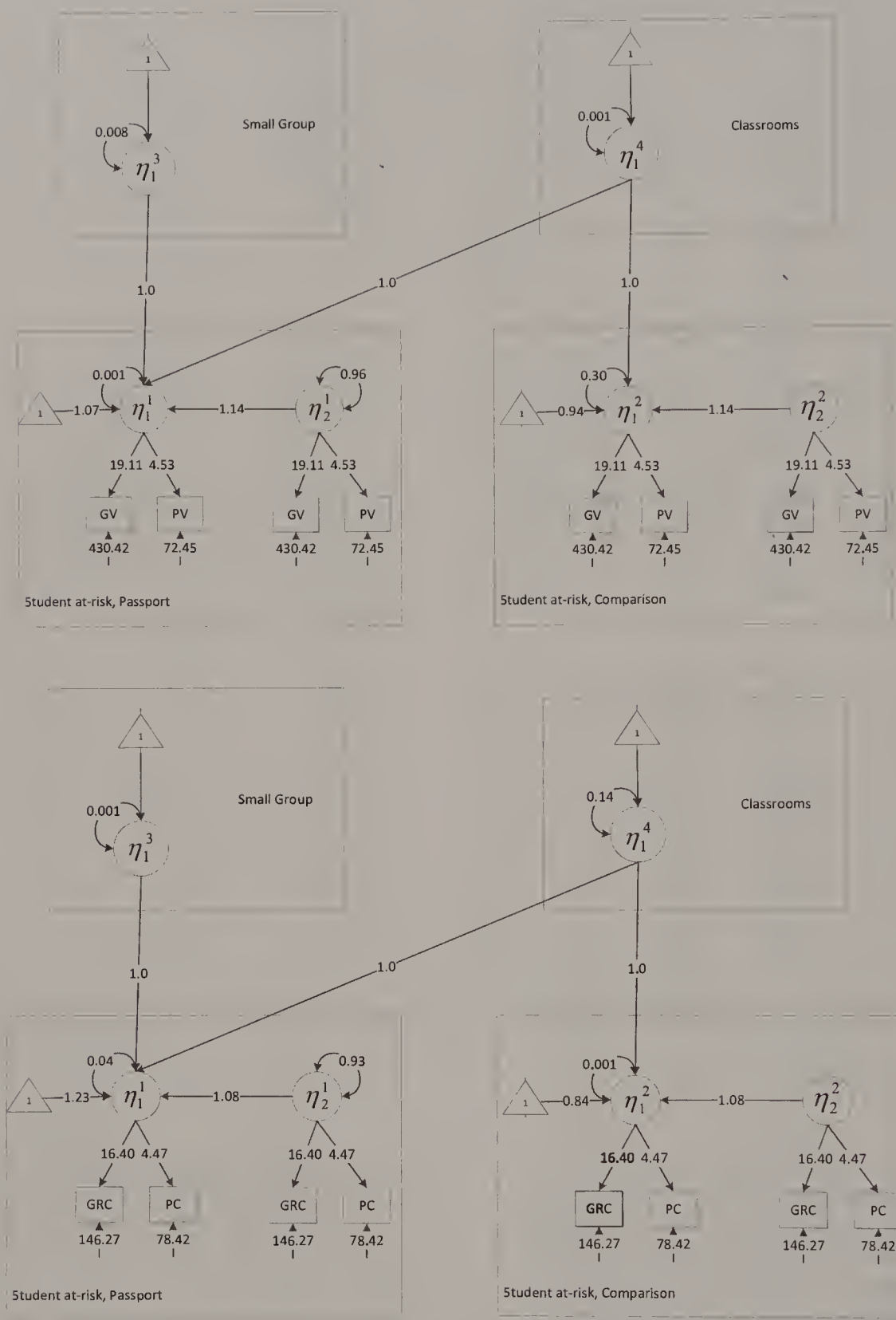


Figure 5. Exploratory *n*-level structural equation modeling (SEM) for non-English Learners on vocabulary (top) and reading comprehension (bottom).

the individual measures that comprised our latent variable in the present study (i.e., WJIII passage comprehension [ES = 0.14] and the GMRT [ES = 0.28]). It is noteworthy that 0.38 exceeds the effect size criteria of 0.25 for substantively important impact from the What Works Clearinghouse (2014). On the basis of the mean standard scores, students in the comparison group appeared to make expected progress (1 year's worth of progress) in reading

comprehension, whereas students in the treatment group accelerated their learning. In other words, students in the comparison group didn't fall any further behind whereas students in the treatment group made some progress toward closing the gap between their current level of performance and expected levels of performance in reading comprehension. Importantly, neither group of students demonstrated on grade level performance at the end of the



intervention, although the accelerated learning in reading comprehension for students in the treatment group is promising. We found no significant differences between study groups on word reading or vocabulary. Thus, our findings suggest participation in Passport to Literacy can improve student reading comprehension, which is a finding consistent with our initial work (Wanzek et al., 2016).

That we found no main effects for word reading or vocabulary is important, particularly as it is consistent with our prior study and suggests that for students with weak comprehension, participating in Passport to Literacy would likely move the dial on only on reading comprehension. This is likely because, although the program is multicomponent, it focuses primarily on reading comprehension, with relatively limited word work or in-depth vocabulary instruction. Our observations indicated that, as designed, on average more than 40% of the treatment intervention was devoted to explicit instruction in reading comprehension. In contrast, the percentages of implemented intervention devoted to vocabulary, text reading, decoding, and spelling were 21%, 17%, 12%, and 5%, respectively. The quality for this instruction was fairly high as well, indicating students received explicit, systematic instruction in reading comprehension. This high-quality, comprehension emphasis in the intervention may explain the reading comprehension outcomes students realized. In other words, the fact that Passport to Literacy has its benefits largely in the area of reading comprehension may be related to the focus of the intervention. The effect sizes for reading comprehension in the present study are larger than those in our prior study (effect sizes ranged from 0.14 to 0.28 in the prior study), but are smaller than effect sizes reported in two other multicomponent interventions. Specifically, for reading comprehension measures, Vadasy and Sanders (2008) reported an effect size of 0.50 and O'Connor et al.'s (2002) effect sizes ranged from 1.39 to 1.46. By contrast, Ritchey et al. (2012) found no significant differences on a standardized measure of reading comprehension, but did report an effect size of 0.56 on a researcher-developed measure of comprehension strategy use.

In our previous study of the effects of Passport to Literacy with a smaller sample, we suggested that our pattern of effects (significant effects for reading comprehension, but not for word reading or vocabulary) might be related to the amount of time attributed to narrative and expository comprehension and word reading during the lessons, with an average of 12 min of reading comprehension instruction and 6 min of vocabulary instruction in a typical half hour lesson, compared with 3 min of decoding or word reading instruction. In contrast, the interventions in the O'Connor et al. (2002) and Vadasy and Sanders (2008) studies included relatively more fluency practice than in the current study, perhaps allowing students to access greater amounts of text for improving their overall reading comprehension. The samples in the studies by O'Connor et al. (2002) as well as Vadasy and Sanders presented with lower overall word recognition and fluency abilities initially as well. Ritchey et al. (2012) emphasized fluency and expository comprehension, but for a briefer period of time (24 sessions) than O'Connor et al. (2002), Vadasy and Sanders, or the current study. The brief time period makes it difficult to directly compare the relationship between the instruction in the intervention and findings to these other more lengthy studies. However, the current findings seem to align with the differences in intervention focus, length of intervention, and results of the previous studies.

Our second research question addressed moderation, to help inform for whom the intervention was effective. We hypothesized, based on exploratory findings from our previous study, that students with low levels of initial comprehension might demonstrate less growth than students with better initial comprehension. However, with our larger sample and using latent variables, we found no moderation effects for initial status on comprehension, suggesting the intervention was equally beneficial for students at all levels of initial comprehension. This is encouraging for practice as the intervention, with its relative emphasis on comprehension, can assist all levels of struggling, upper elementary students in improving their reading comprehension. There was also no moderation of the intervention effects for reading comprehension on the basis of students' initial reading fluency, a finding that aligns with O'Connor et al. (2002), though O'Connor et al. categorized students into lower or higher fluency students based on a break point. We examined moderation of oral reading fluency differences as a continuous variable. The intervention was equally beneficial in improving reading comprehension for students at all levels of initial reading fluency. However, we did find that initial individual differences in word reading ability significantly moderated the effect of the treatment, with students who entered the intervention at lower levels of word recognition making less progress in reading comprehension than students who entered the intervention with higher levels of word reading. An implication for schools is that these students with low word reading may require a reading intervention that incorporates more word study before they can fully benefit from an intervention that emphasizes reading comprehension. The relatively brief intensive word study provided at the beginning of the Passport to Literacy intervention may not be enough for students with low word recognition to make the same gains as those entering with higher levels of word recognition. Torgesen et al. (2001) implemented an intensive reading intervention largely focused on word recognition for students with very low initial word reading skills and reported significant gains in standard scores across word reading and reading comprehension. The lack of control group in the Torgesen et al. study makes it difficult to compare effect sizes with other studies, but an intensive intervention with a heavier emphasis on word recognition is likely needed for students with the lowest word recognition abilities at the upper elementary level. To summarize, the Passport to Literacy intervention provided improvements in students' reading comprehension beyond the typical school services for students at varying levels of initial reading comprehension or reading fluency but who had relatively higher levels of word reading ability.

Encouragingly, the effects of the intervention on reading comprehension were similar for EL and non-EL students ( $ES = 0.38$  and  $0.39$ , respectively), suggesting the intervention is equally beneficial and appropriate for ELs to improve their reading and understanding of English text. Practical benefits of the intervention were noted in relation to word reading for the EL students, but this was not a significant moderation. Previous work reviewed by Baker et al. (2014) demonstrated that both younger ELs (kindergarten through Grade 1) and older ELs (Grades 6 through 8) benefit from small group multicomponent reading interventions in terms of word reading and comprehension. Wanzek and Roberts (2012) also noted EL status moderated effects on word attack and word identification with the EL students performing better than non-EL students following intervention. These higher effects oc-



curred regardless of the emphasis of the intervention (e.g., comprehension emphasis, word recognition emphasis).

## Limitations

Although our study was rigorous, there are always limitations involved with school-based research. To ensure a strong test of the efficacy of the Passport to Literacy intervention, we trained research staff to implement the intervention with a high degree of fidelity and dosage consistent with the publisher's recommendations. Thus, similar effects may or may not be achieved by school personnel depending on implementation. We also recruited schools that were diverse and served students from low socioeconomic backgrounds, so our findings might not generalize to schools serving students from higher socioeconomic backgrounds. The majority of our ELs in our study were Hispanic and our findings may not generalize to students from other language backgrounds, particularly those with orthographies that are very different than English. Further, effect sizes are interpretable relative to the comparison condition in the participating schools where very few struggling readers received supplemental interventions as a part of their typical practice.

## Implications and Directions for Future Research

Teachers and school leaders face challenges in identifying effective reading interventions for students in the upper elementary grades, particularly given the high numbers of students who continue to struggle with reading after third grade (National Center for Educational Statistics, 2016). The increased demands placed on students beginning in fourth grade may cause a slowing of actual versus expected growth for some students (Chall & Jacobs, 1983). Therefore, fourth grade teachers are often faced with the challenge of providing intervention not only for students with previously identified reading difficulties that have not been adequately remediated, but also students with late-emerging reading difficulties (Compton, Fuchs, Fuchs, Elleman, & Gilbert, 2008).

The current study suggests that a multicomponent intervention emphasizing comprehension instruction can allow students to accelerate their reading comprehension outcomes. Without such interventions, particularly given the limited emphasis within core classroom instruction to support learning to read in fourth grade, students who do not read proficiently could face serious and ongoing consequences, not only in reading language arts, but also across content areas.

On the one hand, the positive effects for reading comprehension found in our study extend the limited evidence base on effective multicomponent reading interventions for upper elementary students. On the other hand, the lack of effects for word reading or vocabulary underscores the need for more research on intensive interventions for fourth grade students with the most severe reading difficulties. For example, there is an even more intensive level of the Passport to Literacy intervention, which the publisher recommends for students in need of more intensive levels of instruction. It is more intensive in that students are served in smaller groups and for a longer session and includes additional instruction, including instruction in reading fluency that has been more emphasized in previous work (O'Connor et al., 2002; Vadasy & Sanders, 2008). It is possible that this extended intervention will be

more potent than the standard implementation of the Passport to Literacy intervention, providing the additional emphasis without decreasing the time spent on comprehension. To guide schools' intervention implementation for the upper elementary grades, additional research is needed to identify appropriate interventions, describe for whom they are effective, and to examine the relative benefits of interventions with increasing intensity to meet adequately meet the varying needs of students.

## References

- Baker, S., Lesaux, N., Jayanthi, M., Dimino, J., Proctor, C. P., Morris, J., . . . Newman-Gonchar, R. (2014). *Teaching academic content and literacy to English learners in elementary and middle school* (NCEE 2014–4012). Washington, DC: U.S. Department of Education.
- Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods, 16*, 149–165. <http://dx.doi.org/10.1037/a0023464>
- Blachman, B. A., Schatschneider, C., Fletcher, J. M., Francis, D. J., Clonan, S. M., Shaywitz, B. A., & Shaywitz, S. E. (2004). Effects of intensive reading remediation for second and third graders and a 1-year follow-up. *Journal of Educational Psychology, 96*, 444–461. <http://dx.doi.org/10.1037/0022-0663.96.3.444>
- Chall, J. S., & Jacobs, V. A. (1983). Writing and reading in the elementary grades: Developmental trends among low-SES children. *Language Arts, 60*, 617–626.
- Compton, D. L., Fuchs, D., Fuchs, L. S., Elleman, A. M., & Gilbert, J. K. (2008). Tracking children who fly below the radar: Latent transition modeling of students with late-emerging reading disability. *Learning and Individual Differences, 18*, 329–337. <http://dx.doi.org/10.1016/j.lindif.2008.04.003>
- Edmonds, M., & Briggs, K. (2003). The instructional content emphasis instrument: Observations of reading instruction. In S. Vaughn & K. L. Briggs (Eds.), *Reading in the classroom: Systems for the observation of teaching and learning* (pp. 31–52). Baltimore, MD: Brookes/Cole.
- Flavell, J. H. (1992). Cognitive development: Past, present, and future. *Developmental Psychology, 28*, 998–1005. <http://dx.doi.org/10.1037/0012-1649.28.6.998>
- Garvan, C. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology, 91*, 579–593. <http://dx.doi.org/10.1037/0022-0663.91.4.579>
- Goddard, R., Goddard, Y., Kim, E. S., & Miller, R. (2015). A theoretical and empirical analysis of the roles of instructional leadership, teacher collaboration, and collective efficacy beliefs in support of student learning. *American Journal of Education, 121*, 501–530. <http://dx.doi.org/10.1086/681925>
- Good, R. H., & Kaminski, R. (2002). *Dynamic Indicators of Basic Early Literacy Skills 6th Edition (DIBELS)*. Eugene, OR: Institute for the Development of Educational Achievement. Retrieved from <http://dibels.uoregon.edu/>
- Good, R. H., Kaminski, R. A., Shinn, M., Bratten, J., Shinn, M., Laimon, D., . . . Flindt, N. (2004). *Technical adequacy of DIBELS: Results of the early childhood research institute on measuring growth and development* (Tech. Rep. No., No. 7). Eugene, OR: University of Oregon.
- Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*. New York, NY: Routledge.
- Kamil, M. L., Borman, G. D., Dole, J., Kral, C. C., Salinger, T., & Torgesen, J. (2008). *Improving adolescent literacy: Effective classroom and intervention practices. IES practice guide (NCEE 2008–4027)*. Washington, DC: National Center for Education Evaluation and Regional Assistance.



- Kieffer, M. J. (2008). Catching up or falling behind? Initial English proficiency, concentrated poverty, and the reading growth of language minority learners in the United States. *Journal of Educational Psychology, 100*, 851–868. <http://dx.doi.org/10.1037/0022-0663.100.4.851>
- Kieffer, M. J. (2010). Socioeconomic status, English proficiency, and late-emerging reading difficulties. *Educational Researcher, 39*, 484–486. <http://dx.doi.org/10.3102/0013189X10378400>
- Kieffer, M. J. (2014). Morphological awareness and reading difficulties in adolescent Spanish-speaking language minority learners and their classmates. *Journal of Learning Disabilities, 47*, 44–53. <http://dx.doi.org/10.1177/0022219413509968>
- Lohr, S., Schochet, P. Z., & Sanders, E. (2014). *Partially nested randomized controlled trials in education research: A guide to design and analysis (NCER 2014–2000)*. Washington, DC: U.S. Department of Education. Retrieved from <http://ies.ed.gov/ncerp/pubs/20142000/pdf/20142000.pdf/>
- MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2006). *Gates-MacGinitie Reading Tests* (4th ed.). Rolling Meadows, IL: Riverside Publishing.
- Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly, 40*, 148–182. <http://dx.doi.org/10.1598/RRQ.40.2.2>
- Mason, L. H. (2004). Explicit self-regulated strategy development versus reciprocal questioning: Effects on expository reading comprehension among struggling readers. *Journal of Educational Psychology, 96*, 283–296. <http://dx.doi.org/10.1037/0022-0663.96.2.283>
- Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*, 259–284. <http://dx.doi.org/10.1037/1082-989X.10.3.259>
- Miranda, A., Villaescusa, M. I., & Vidal-Abarca, E. (1997). Is attribution retraining necessary? Use of self-regulation procedures for enhancing the reading comprehension strategies of children with learning disabilities. *Journal of Learning Disabilities, 30*, 503–512. <http://dx.doi.org/10.1177/002221949703000506>
- National Center for Educational Statistics. (2016). *National assessment of educational progress: The nation's report card*. Washington, DC: U.S. Department of Education.
- O'Connor, R. E., Bell, K. M., Harty, K. R., Larkin, L. K., Sackor, S. M., & Zigmond, N. (2002). Teaching reading to poor readers in the intermediate grades: A comparison of text difficulty. *Journal of Educational Psychology, 94*, 474–485. <http://dx.doi.org/10.1037/0022-0663.94.3.474>
- O'Connor, R. E., Fulmer, D., Harty, K. R., & Bell, K. M. (2005). Layers of reading intervention in kindergarten through third grade: Changes in teaching and student outcomes. *Journal of Learning Disabilities, 38*, 440–455. <http://dx.doi.org/10.1177/00222194050380050701>
- Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*, 117–175. [http://dx.doi.org/10.1207/s1532690xci0102\\_1](http://dx.doi.org/10.1207/s1532690xci0102_1)
- Ritchey, K. D., Silverman, R. D., Montanaro, E. A., Speece, D. L., & Schatschneider, C. (2012). Effects of a tier 2 supplemental reading intervention for at-risk fourth-grade students. *Exceptional Children, 78*, 318–334. <http://dx.doi.org/10.1177/001440291207800304>
- Scammacca, N., Roberts, G., Vaughn, S., Edmonds, M., Wexler, J., Reutebuch, C. K., & Torgesen, J. K. (2007). *Interventions for adolescent struggling readers: A meta-analysis with implications for practice*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.
- Sterba, S. K., Preacher, K. J., Forehand, R., Hardcastle, E. J., Cole, D. A., & Compas, B. E. (2014). Structural equation modeling approaches for analyzing partially nested data. *Multivariate Behavioral Research, 49*, 93–118. <http://dx.doi.org/10.1080/00273171.2014.882253>
- Swanson, H. L., Hoskyn, M., & Lee, C. (1999). *Interventions for students with learning disabilities: A meta-analysis of treatment outcomes*. New York, NY: Guilford Press.
- Therrien, W. J., Wickstrom, K., & Jones, K. (2006). Effect of a combined repeated reading and question generation intervention on reading achievement. *Learning Disabilities Research & Practice, 21*, 89–97. <http://dx.doi.org/10.1111/j.1540-5826.2006.00209.x>
- Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Rose, E., Lindamood, P., Conway, T., et al. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology, 91*, 579–593. <http://dx.doi.org/10.1037/0022-0663.91.4.579>
- Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34*, 33–58. <http://dx.doi.org/10.1177/002221940103400104>
- Torgesen, J. K., Houston, D. D., Rissman, L. M., Decker, S. M., Roberts, G., Vaughn, S., . . . Lesaux, N. (2007). *Academic literacy instruction for adolescents: A guidance document from the Center on Instruction*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.
- Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Lindamood, P., Rose, E., Conway, T., . . . Sanders, E. A. (2008). Repeated reading intervention: Outcomes and interactions with readers' skills and classroom instruction. *Journal of Educational Psychology, 100*, 272–290. <http://dx.doi.org/10.1037/0022-0663.100.2.272>
- Vadasy, P. F., & Sanders, E. A. (2008). Repeated reading intervention: Outcomes and interactions with readers' skills and classroom instruction. *Journal of Educational Psychology, 100*, 272–290. <http://dx.doi.org/10.1037/0022-0663.100.2.272>
- Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R., & Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88*, 601–638. <http://dx.doi.org/10.1037/0022-0663.88.4.601>
- Wanzek, J., Petscher, Y., Al Otaiba, S., Kent, S. C., Schatschneider, C., Haynes, M., & Jones, F. (2016). Examining the average and local effects of a standardized treatment for fourth graders with reading difficulties. *Journal of Research on Educational Effectiveness, 9*, 45–66. <http://dx.doi.org/10.1080/19345747.2015.1116032>
- Wanzek, J., & Roberts, G. (2012). Reading interventions with varying instructional emphases for fourth graders with reading difficulties. *Learning Disability Quarterly, 35*, 90–101. <http://dx.doi.org/10.1177/0731948711434047>
- Wanzek, J., Wexler, J., Vaughn, S., & Ciullo, S. (2010). Reading interventions for struggling readers in the upper elementary grades: A synthesis of 20 years of research. *Reading and Writing, 23*, 889–912. <http://dx.doi.org/10.1007/s11145-009-9179-5>
- What Works Clearinghouse. (2014). *Procedures and standards handbook (Version 3.0)*. Washington, DC: U.S. Department of Education. Retrieved from [http://ies.ed.gov/ncee/wwc/pdf/reference\\_resources/wwc\\_procedures\\_v3\\_0\\_standards\\_handbook.pdf](http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf)
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement*. Itasca, IL: Riverside.

Received August 15, 2016

Revision received December 5, 2016

Accepted January 7, 2017 ■

# Examining the Relations Between Executive Function, Math, and Literacy During the Transition to Kindergarten: A Multi-Analytic Approach

Sara A. Schmitt  
Purdue University

G. John Geldhof  
Oregon State University

David J. Purpura  
Purdue University

Robert Duncan and Megan M. McClelland  
Oregon State University

The present study explored the bidirectional and longitudinal associations between executive function (EF) and early academic skills (math and literacy) across 4 waves of measurement during the transition from preschool to kindergarten using 2 complementary analytical approaches: cross-lagged panel modeling and latent growth curve modeling (LGCM). Participants included 424 children (49% female). On average, children were approximately 4.5 years old at the beginning of the study ( $M = 4.69$ ,  $SD = .30$ ) and 55% were enrolled in Head Start. Cross-lagged panel models indicated bidirectional relations between EF and math over preschool, which became directional in kindergarten with only EF predicting math. Moreover, there was a bidirectional relation between math and literacy that emerged in kindergarten. Similarly, LGCM revealed correlated growth between EF and math as well as math and literacy, but not EF and literacy. Exploring the patterns of relations across the waves of the panel model in conjunction with the patterns of relations between intercepts and slopes in the LGCMs led to a more nuanced understanding of the relations between EF and academic skills across preschool and kindergarten. Implications for future research on instruction and intervention development are discussed.

**Keywords:** executive function, mathematics, literacy, preschool, kindergarten

Over the past decade, there has been increased focus on children's executive function (EF)—specifically on its development and how it relates to other school readiness domains. One reason for this surge of interest is that EF in early childhood has been connected to a range of critical developmental outcomes, including physical health, social-emotional well-being, and occupational attainment in adulthood (Moffitt et al., 2011). Of particular interest are the significant and direct relations found between early EF and academic achievement. Findings from a number of studies indicate that individual differences in EF measured in early childhood predict concurrent and long-term math and literacy achievement (Duckworth, Tsukayama, & May, 2010; Fuhs, Nesbitt, Farran, & Dong, 2014; McClelland, Acock, Piccinin, Rhea, & Stallings,

2013) as well as growth in children's higher-level reasoning strategies (Richland & Burchinal, 2013).

Although the predictive link between EF and early achievement is established, it is less clear whether early academic skills also predict the development of EF. Recent evidence indicates that there may be a bidirectional association between EF and academic skills, particularly for math (Fuhs et al., 2014; Welsh, Nix, Blair, Bierman, & Nelson, 2010). However, these studies were limited to just three time points over the course of the preschool and kindergarten years and by the analytic approach employed (i.e., only panel models were used). Further, it is unclear whether growth trajectories in EF are related to growth trajectories in other domains (e.g., math). The overarching goal of the current study was therefore to examine the longitudinal relations between EF and academic skills across four waves of measurement during the transition from preschool to kindergarten. We had two specific aims. First, we investigated the bidirectional relations between EF and academic skills (math and literacy) through a longitudinal panel model that tested whether relative standing on the domains at each time point was related to changes in relative standing on the other domains. Second, we examined relations between growth in EF, math, and literacy using latent growth curve models (LGCM) that tested whether the rate of absolute change across all time points on the domains was correlated. The two models provide unique information by identifying when early skills are most related to subsequent skill development (i.e., panel models), and to what extent children's overall growth on skills during this developmental period are related (i.e., LGCM).

---

This article was published Online First March 23, 2017.

Sara A. Schmitt, Department of Human Development and Family Studies, Purdue University; G. John Geldhof, Human Development and Family Sciences and Hallie E. Ford Center for Healthy Children and Families, Oregon State University; David J. Purpura, Department of Human Development and Family Studies, Purdue University; Robert Duncan and Megan M. McClelland, Human Development and Family Sciences and Hallie E. Ford Center for Healthy Children and Families, Oregon State University.

This study was supported by Grant R305A100566 from the U.S. Department of Education Institute for Education Sciences.

Correspondence concerning this article should be addressed to Sara A. Schmitt, 1202 West State Street, West Lafayette, IN 47907. E-mail: saraschmitt@purdue.edu



### Importance of EF for Academic Achievement

EF emerges early in life and develops across the life span; however, structural changes in the prefrontal cortex between ages two and five allow for dramatic increases in EF skills during early childhood (Zelazo & Ulrich, 2011). Evidence suggests that EF involves three related, yet distinct, cognitive processes (Miyake et al., 2000): working memory (holding information in mind while processing other information; Gathercole, Pickering, Knight, & Stegmann, 2004), inhibitory control (overriding a dominant response; Dowsett & Livesey, 2000), and cognitive flexibility or attention shifting (maintaining focus and flexibly adapting to changing goals; Rueda, Posner, & Rothbart, 2005). When children enter kindergarten, they must adapt to new, more formal, and structured educational contexts that may require greater EF to navigate, compared to the less formal and structured educational environments experienced earlier.

The transition from preschool to kindergarten is not only an important developmental period for EF, it is also a time when early academic skills develop rapidly. Similar to EF, a substantial body of research highlights the importance of the preschool years for the development of early literacy (e.g., National Early Literacy Panel, 2008; Whitehurst & Lonigan, 1998) and math skills (e.g., National Mathematics Advisory Panel, 2008), and it is well known that early academic skills are precursors to later academic success (e.g., La Paro & Pianta, 2000; Stevenson & Newman, 1986). Furthermore, evidence supports an association between early math and reading (Duncan et al., 2007; LeFevre et al., 2010; Purpura, Hume, Sims, & Lonigan, 2011). These two academic domains are related over time and children who demonstrate difficulties in one area are at elevated risk for having difficulties in the other (Barbarisi, Katusic, Colligan, Weaver, & Jacobsen, 2005). Theory and research suggest that aspects of literacy may be foundational for math development. Children may need to draw upon vocabulary skills to learn number words and complete math tasks that are inherently language based (LeFevre et al., 2010; Purpura et al., 2011). Although EF, early math, and emergent literacy appear to develop during the same time frame, some scholars argue that EF is foundational for academic achievement (Blair & Raver, 2015; McClelland et al., 2007). Furthermore, children's EF is related to both their own and their peers' acquisition of academic skills (Skibbe, Phillips, Day, Brophy-Herb, & Connor, 2012). For example, Skibbe and colleagues (2012) found that children demonstrated greater gains in literacy skills during the academic year when they were part of classrooms where their classmates had higher levels of EF.

Theoretical and empirical perspectives support the connection between EF and math and literacy skills. For children to take advantage of learning opportunities in classroom contexts, they must be able to pay attention, persist on challenging tasks, and avoid distractions (Blair & Raver, 2015; McClelland, Geldhof, Cameron, & Wanless, 2015). Specifically, strong EF may be critical for aspects of early math development such as cardinality or formal addition, which require children to flexibly shift attention from procedural to more conceptual problem elements and inhibit previously learned rules. Similarly, EF may be needed for growth in emergent literacy skills, such as phonological awareness, because children must have the ability to hold letter sounds in mind and switch between combining and separating sounds and

words. Studies suggest there is a predictive relation between EF and math and literacy achievement in diverse samples of young children, even after controlling for relevant sociodemographic factors (e.g., maternal education, child IQ) and initial achievement scores (Bull, Espy, & Wiebe, 2008; Duncan et al., 2007; McClelland, Acock, & Morrison, 2006). Findings from a recent study demonstrate a long-term relation between EF and achievement, such that children who were rated higher on aspects of EF (e.g., attention and persistence) during preschool were more likely to complete college (McClelland et al., 2013). Even among children with academic difficulties (i.e., those who experienced grade retention), EF appears to play a role in subsequent math and reading growth. For example, Chen, Hughes, and Kwok (2014) found that, among children who had been held back a grade, those who exhibited patterns of more rapid academic growth displayed higher EF skills.

Although prior evidence suggests EF is associated with both math and literacy in early childhood, the concurrent and predictive relation between EF and math seems to be stronger than the relation between EF and literacy in young children (Blair & Razza, 2007; Blair, Ursache, Greenberg, & Vernon-Feagans, 2015; Cameron Ponitz, McClelland, Matthews, & Morrison, 2009; Schmitt, Pratt, & McClelland, 2014). Furthermore, EF skills may mediate the development of math skills across the early elementary years, but not the development of literacy skills (Hassinger-Das, Jordan, Glutting, Irwin, & Dyson, 2014). Several interpretations explaining these differential associations have been introduced in recent literature. For example, one interpretation is that math content places more cognitive demands on children than does literacy content. Math skills, therefore, may require stronger EF skills to acquire (Bull et al., 2008; Clark, Pritchard, & Woodward, 2010; Espy et al., 2004; Willoughby, Blair, Wirth, & Greenberg, 2012). Evidence from the neuroscience literature also indicates an overlap between the brain regions that support EF and math development (Klingberg, 2006), suggesting that growth in EF may strongly facilitate growth in math while having a weaker influence on the development of literacy. A second interpretation is that this relation results from instructional content (or lack thereof) provided in early childhood classrooms (Fuhs et al., 2014). Preschool teachers spend significantly more time engaged in direct literacy instruction than in math instruction, suggesting that children may need to seek out their own independent math activities which may be influenced by their EF skills. For example, children who have stronger levels of EF may choose more complex and difficult math activities during free play (or may be directed to by teachers and parents) because they may be more cognitively ready to do so. A third interpretation is that EF provides a foundation for the development of reasoning abilities or fluid mental capacities (e.g., problem solving), which are typically required to do well on many math assessments (Blair et al., 2015). In contrast, many literacy assessments are more knowledge-based, making stronger demands on crystallized mental abilities (e.g., vocabulary) and fewer demands on EF and fluid mental abilities.

### Bidirectional Relations Between EF and Academic Skills

Although EF is considered by some to be foundational for the development of academic skills, recent analyses have the bidi-



rectionality between EF and achievement (Fuhs et al., 2014; Welsh et al., 2010). Indeed, early academic skills may be important for the development of EF, just as EF is important for the development of early academic skills. Although the ability to pay attention, remember complex rules, and persist on challenging tasks likely helps children perform better academically (Blair et al., 2007; Blair & Raver, 2015), strong academic skills may also contribute to children's ability to sustain attention, remember a series of rules, and inhibit incorrect responses on complex tasks (Fuhs et al., 2014). Engaging in a complex math activity, for example, requires children to identify the quantities of multiple sets, retain those quantities in memory, and compare them.

Recent empirical evidence has suggested that there may be a bidirectional relation between direct assessments of EF and academic skills. In one study assessing developmental associations between EF and academic skills during the prekindergarten year, EF at the beginning of the year predicted gains in math and literacy; however, math at the beginning of prekindergarten also predicted gains in EF (Welsh et al., 2010). In a second study, Fuhs and colleagues (2014) found reciprocal associations between EF and math. These associations were maintained across preschool, and, although EF continued to predict math through kindergarten, the predictive relation of math on EF dissipated between the end of preschool and end of kindergarten. However, it was not clear when during this year the predictive association of math on EF faded. Results from Fuhs and colleagues' (2014) also indicated a reciprocal relation between EF and oral comprehension skills across the prekindergarten year, but not for other literacy skills. These findings provide initial evidence for a bidirectional relation between EF and early achievement; however, the analyses utilized in these studies were limited to three time points (fall and spring of preschool and spring of kindergarten). The addition of a fourth time point at the beginning of kindergarten is needed to understand these relations more thoroughly; significant changes in children's experiences and in the development of EF and academic skills may occur between the spring of preschool and the spring of kindergarten. Further, the addition of a fourth time point allows us to determine when early EF, math, and literacy interventions may be most beneficial and likely to facilitate cross-domain growth. Identifying more specific and precise times at which these relations may change could have applied implications as children enter and move through kindergarten. Moreover, the relations between these variables at different ages remain unclear.

### Correlated Growth Between EF and Academic Skills

In addition to a need for more research on the bidirectional relations between EF and academic skills, there is a dearth in extant literature exploring whether the rates of change in these domains are correlated during the transition to kindergarten. Understanding whether growth in one domain is related to growth in another has theoretical as well as practical implications for instruction and intervention. From a theoretical standpoint, exploring correlated growth across domains will help us understand the potential that improvements in one domain lead to improvements in another domain or that other individual or environmental factors may be influencing EF, math, and literacy development similarly over time (Willoughby, Kupersmidt, & Voegler-Lee, 2012), rather

than earlier skills in and of themselves. Some have suggested that the relation between EF and math may be attributable to other factors such as IQ, but there is also evidence showing that EF is separate from IQ (e.g., Blair, 2006). From a practical standpoint, if improvements in EF are associated with improvements in math, instruction or intervention efforts focused on EF may also have a beneficial effect on children's math development. Likewise, instruction or intervention efforts focused on math or literacy could have beneficial effects on children's EF development. For example, engaging in math activities may not only support the development of math concepts, but doing so may also allow children to practice EF skills (e.g., attending to details, remembering instructions). Similarly, retaining details of a story in memory while simultaneously attending to new developments in the plotline to comprehend the broader story also may provide children an opportunity to practice EF skills.

To our knowledge, no studies to date have examined dual trajectory latent growth curves between EF and academic skills. In one related study, fixed effects models were used to explore whether intraindividual change on measures of EF predicted intraindividual change in math, literacy, and vocabulary during the transition to kindergarten. Results indicated that growth in EF on some, but not all, of the measures predicted growth in math, and that growth on one measure of inhibitory control was related to vocabulary development (McClelland et al., 2014). The current study extends these analyses by using LGCM. Although using fixed effects models can be informative, this type of analysis only explores relations between intraindividual changes over time between the domains. In the current study, we attempted to more accurately measure children's trajectories using LGCM, which estimates associations across domains on random intercepts and linear and quadratic slopes. Further, LGCM is able to estimate the associations between the EF, math, and literacy slopes, conditional on differences in their intercepts. However, the LGCM is not able to determine whether one domain contributes to or is causally related to development in another, or whether other factors simultaneously influence multiple domains of development (e.g., high quality early math instruction). Once correlated growth is established, follow-up studies would be needed to further elucidate the relations between cross-domain growth trajectories.

### Multi-Analytic Approach

Previous work exploring longitudinal relations between EF and early academic skills has typically taken a single-analysis approach, and this approach has primarily been panel models (e.g., path analysis). Although findings from single-analysis studies have been useful, they provide limited information on the development of these important skills. As Greene and colleagues noted more than two decades ago, "all methods have inherent biases and limitations, so use of only one method to assess a given phenomenon will inevitably yield biased and limited results" (Greene, Caracelli, & Graham, 1989, p.256; see also Campbell & Fiske, 1959; Symonds & Gorard, 2010). Thus, we took a multi-analytic approach to address our overarching research goal: examining the longitudinal associations between EF and achievement. We first implemented a cross-lagged panel model using a latent EF factor to determine how children's relative standing on measures of EF, math, and literacy was related over time. That is, stability and



cross-lagged effects in cross-lagged panel models determine the stability of participants' relative standing on a variable without regard for whether the sample (or individual participants) actually exhibited gains in absolute magnitude. High stability indicates that participants who scored higher than the sample mean at one time point tend to score higher on the sample mean at the previous time point, regardless of whether that sample mean increased, decreased, or remained the same (see also Wu, Selig, & Little, 2013).

Previous work examining the bidirectional associations between EF and academic skills using cross-lagged panel models (e.g., Fuhs et al., 2014) has relied on factor scores rather than modeling latent associations directly. The use of factor scores as dependent variables is known to produce biased regression slopes and standard errors (Muthén, 2011; Skrondal & Laake, 2001). The extent that previous findings are biased by a reliance on factor scores therefore remains unclear, and we overcome this limitation by modeling EF directly as a latent factor.

We also addressed our research goal using a series of LGCMs that examined absolute changes (i.e., sample- and individual-level growth) in EF, math, and literacy. By examining absolute changes, the LGCMs allowed us to paint a more complete picture of how EF, math, and literacy codevelop by demonstrating to what extent growth in one domain is related to growth in another domain during the same time frame. Thus, the panel models allowed us to examine the bidirectional relations between EF and achievement and whether relative standing on one domain predicts relative standing on another domain at the subsequent time point, and the LGCMs allowed us to examine changes in absolute magnitude and relations between growth trajectories across all four time points (Wu et al., 2013).

### The Present Study

The goal of the present study was to clarify and expand upon prior work (e.g., Fuhs et al., 2014; McClelland et al., 2007) that has examined the longitudinal relations between EF, math, and literacy across the transition to kindergarten (preschool-kindergarten). More specifically, we aimed to paint a broader picture of how EF, math and literacy are associated over time. Based on recent theoretical and empirical evidence indicating that EF and math may be tightly coupled constructs and reciprocally related (Fuhs et al., 2014; McClelland et al., 2015), we hypothesized that EF would significantly predict math in preschool and kindergarten, and also that math would predict math during this time frame. Further, we expected that EF and math growth trajectories would be correlated, although previous research on associations between intraindividual change between the two domains is mixed (McClelland et al., 2014; Willoughby, Kupersmidt, & Voegler-Lee, 2012). Previous research has shown inconsistent links between EF and literacy (Blair & Razza, 2007; Cameron Ponitz et al., 2009; Schmitt et al., 2014), nonsignificant bidirectional associations (Fuhs et al., 2014), and nonsignificant associations for intraindividual change models (McClelland et al., 2014). We therefore did not expect that this same reciprocal association would emerge for EF and literacy, nor did we expect the EF and literacy growth trajectories to be correlated. We also hypothesized a bidirectional relation as well as correlated growth between math and literacy due to the noted strong relation between early math and literacy skills over the

preschool and kindergarten years (Duncan et al., 2007; LeFevre et al., 2010; Purpura et al., 2011).

Findings from this study will contribute to the existing literature in multiple ways. In contrast to other studies, we have four data points (fall and spring of preschool and kindergarten), which will allow us to explore changes in the relations as well as growth trajectories between these skills during the school year and at critical junctures throughout the transition from preschool to kindergarten at a more fine-grained level. This could have important practical implications for children as they enter and progress through kindergarten. Other studies examining bidirectional associations between EF and early achievement (e.g., Fuhs et al., 2014) were limited to just one data point in kindergarten (end of the year). This additional time point is important in extending previous work because it allows us to better identify at which point between the end of preschool and end of kindergarten the relations between EF and math may change. That is, modeling change in relations across four waves of data collection will allow a better understanding of whether change occurs primarily during the school year (i.e., between Times 1 and 2 and between Times 3 and 4), or if the change is relatively constant across time. In addition, we modeled latent associations directly rather than relying on factor scores that may produce biased results. Finally, no studies to date have examined whether growth trajectories of EF, math, and literacy are correlated using LGCM. Our multi-analytic approach also allows us to examine the same overarching research question using two types of analyses, allowing us to better distill a single story from multiple models that acknowledge diverse ways development can manifest while reducing methodological biases. Results from the present study will further our understanding of the complexity of the relations between EF, math, and literacy, which could have theoretical implications as well as implications for the design and timing of instruction and intervention efforts in preschool and early elementary school.

### Method

#### Participants and Procedure

Children and families ( $N = 435$ ) were recruited from 38 classrooms in 17 preschools in a small city in the Pacific Northwest to participate in a federally funded study focused on refining and evaluating the Head-Toes-Knees-Shoulders (HTKS) task, a direct assessment of EF, as a screening tool for children ages 4–5. As part of this study, several measures of EF as well as a math and literacy assessment were collected at 4 data points from 2011 to 2014. There was no intervention included as part of the larger study that would influence the interpretation of our results. To recruit schools, the principal investigator contacted preschool directors via telephone, e-mail and via individual meetings to invite preschools to be a part of the study. Preschools were selected using a convenience sampling approach (i.e., preschools that were accessible and willing to participate in the study). Children were excluded if they were younger than 4 years old ( $n = 5$ ) or older than 5.5 years old ( $n = 1$ ) in the fall of preschool. Additionally, children were excluded if they did not participate in the study in the fall of preschool ( $n = 5$ ). The remaining 424 eligible children were included in the sample in the current study.



Parents signed a written informed consent statement to allow their child to participate in the study that was approved by the university Institutional Review Board. Children gave verbal assent prior to participating in direct assessments. After consenting to the study, children were assessed in two to three sessions (lasting 10 to 15 min each) during the fall and spring of their preschool and kindergarten years (4 waves total). At each wave of data collection, families received a \$20 gift card for their participation. In the fall of preschool, 55% of the children were enrolled in Head Start and 15% were primarily Spanish speakers (all Spanish speakers were enrolled in Head Start). Teachers identified which children in their classrooms were Spanish-speaking and should receive the assessments in Spanish. We chose this method for identifying Spanish speakers because teachers have the most experience with children in their classroom context and to avoid overtesting children by administering assessments in both languages. Parent demographic questionnaires were collected during the first wave of the study (in Spanish when applicable;  $n = 372$ , 88% response rate). The sample was predominantly reported as White (63%), followed by Latino/Hispanic (19%), multiracial (13%), Asian/Pacific Islander (3%), and other ethnicities (2%). Self-reported parent (87% maternal) education ranged from 0 to 30 years, with an average of approximately two years in college ( $M = 14.40$ ,  $SD = 3.68$ ). Children enrolled in Head Start had parents with significantly lower reported years of education ( $M = 11.58$ ,  $SD = 3.06$ ) than the parents of children not enrolled in Head Start ( $M = 17.34$ ,  $SD = 3.14$ ;  $t(351) = 17.48$ ,  $p < .001$ ). Among children enrolled in Head Start, the primarily Spanish speaking children had parents with significantly lower reported years of education ( $M = 9.08$ ,  $SD = 3.12$ ) than their English-speaking peers ( $M = 12.59$ ,  $SD = 2.38$ ;  $t(178) = 8.17$ ,  $p < .001$ ).

## Measures

At each wave of the study, children were assessed on executive function (EF), literacy, and math skills. EF was assessed with four measures: the Head-Toes-Knees-Shoulders (HTKS) task, a Card Sort task, the Auditory Working Memory subtest from the Woodcock-Johnson III Tests of Cognitive Abilities, and the Simon Says task. Literacy skills were assessed with the Letter-Word Identification subtest from the Woodcock-Johnson III Tests of Achievement Abilities. Math skills were assessed with the Applied Problems subtest from the Woodcock-Johnson III Tests of Achievement Abilities.

**Head-Toes-Knees-Shoulders.** The HTKS was used to assess children's cognitive flexibility, working memory, and inhibitory control through gross motor responses (McClelland & Cameron, 2012; McClelland et al., 2014). In previous research, the measure has been significantly related to measures of cognitive flexibility, working memory, and inhibitory control (see McClelland et al., 2014). There are two parallel forms of the HTKS, which only differ for part one of the assessment (McClelland et al., 2014). The measure includes three sections of 10 items each, with the task becoming progressively harder. In part one, children were instructed to touch their toes (knees in the parallel form) when told to "touch your head (shoulders in the parallel form)" and vice versa. In parts two and three, rules were changed and added, increasing the complexity of the task. Possible scores range from 0 to 60, with a total of 30 test items receiving scores of 0

(incorrect), 1 (self-correct), or 2 (correct). Previous research indicates high interrater agreement ( $\kappa > .90$ ) and evidence supports convergent and predictive validity of this measure when assessing children's EF in culturally diverse samples and in different languages (McClelland et al., 2007, 2014; Wanless, McClelland, Acock, Ponitz, et al., 2011). In the current sample, this measure demonstrated strong internal consistency across all waves (Cronbach's alpha: Wave 1 = .96, Wave 2 = .96, Wave 3 = .96, Wave 4 = .95).

**Card Sort task.** Children's cognitive flexibility was assessed using a Card Sort task similar to the traditional Dimensional Change Card Sort measure (Blackwell, Cepeda, & Munakata, 2009; Frye, Zelazo, & Palfai, 1995; Zelazo, 2006). Administration procedures were similar to those described by Hongwanishkul, Happaney, Lee, and Zelazo (2005). The Card Sorting task consisted of up to 24 items, with each sorting trial having 6 items. During this task, children were asked to sort colored picture cards of a dog, fish, or bird on the basis of three dimensions: color, shape, and size. Four sorting boxes with target cards (either a dog, fish, bird, or frog) affixed on them were placed directly in front of children. The frog target card was meant to be a distractor, and thus, there were no picture cards with frogs on them. The same target and test cards were used for all participants. Children were given one practice trial prior to testing trials. During all test trials, children were given a test card (that had the same picture on it as one of the target cards) and asked the question, "Where does this one go?" and they were to place the card in one of the boxes. No feedback was given. For the first six items (preswitch trial), children were to sort on the basis of shape (e.g., the dog cards go in the sorting box with the dog card affixed). For the second six items (postswitch trial), children were told they were going to play a new game and would now sort on the basis of color. For the third six items (postswitch trial), children were told they were going to play a new game and would now sort on the basis of size. If children scored five or more points on the third section, a fourth set of items were administered which consisted of a new rule: when the card had a black border on it, children were to sort on the basis of size. When the card did not have a black border, children were to sort on the basis of color. All items were weighted equally (including preswitch trial items). Children were given a score of 0 for an incorrect response and 1 for a correct response, with scores ranging from 0 to 24. This assessment demonstrated strong internal consistency in the current sample across all waves for all sections (Cronbach's alpha: Wave 1 = .95, Wave 2 = .93, Wave 3 = .91, Wave 4 = .88).

**Auditory Working Memory.** The Auditory Working Memory subtest from the Woodcock-Johnson III Tests of Cognitive Abilities (Woodcock, McGrew, & Mather, 2001b) or the Bateria III Woodcock-Muñoz (Muñoz-Sandoval, Woodcock, McGrew, & Mather, 2005b) was used to assess children's working memory. The task required children to repeat back to the experimenter things and numbers in a specific order. That is, children had to hold information in mind and then reproduce it in a different order. This standardized task demonstrates strong internal consistency for English-speaking and Spanish-speaking preschool children (Woodcock et al., 2001b). In the current sample, internal consistency was good across all waves for the full sample (Cronbach's alpha: Wave 1 = .87, Wave 2 = .89, Wave 3 = .85, Wave 4 = .82), and for the English-speaking children only (Cronbach's al-



pha: Wave 1 = .89, Wave 2 = .88, Wave 3 = .86, Wave 4 = .81) and the Spanish-speaking children only (Cronbach's alpha: Wave 1 = .92, Wave 2 = .83, Wave 3 = .85, Wave 4 = .91).

**Simon Says task.** The Simon Says task was used to assess inhibitory control (Carlson, 2005; Strommen, 1973). The Simon Says task has been identified in previous research as an advanced anti-imitation task and a measure of inhibitory control in that it requires children to inhibit a prepotent response (i.e., do all requested actions) in favor of a different response (i.e., only do the action if experimenter says "Simon Says"; Carlson, 2005). Specifically, children were asked to perform an action only if the experimenter said, "Simon says," but to remain still otherwise. Of the 10 total trials, five trials required inhibitory control. These trials were scored and children were given a proportion score of the number correct (items requiring inhibitory control). In previous studies, this measure has been significantly correlated with other assessments of inhibitory control (McClelland et al., 2014). Internal consistency for this assessment was good across all waves (Cronbach's alpha: Wave 1 = .87, Wave 2 = .89, Wave 3 = .85, Wave 4 = .82).

**Reliability of EF.** Using the factor loadings presented below and discussed later, we computed composite reliability ( $\omega$ ; McDonald, 1970, 1999; Raykov, 1997; Werts, Linn, & Jöreskog, 1974) for each EF construct.  $\omega$  is identical to Cronbach's (1951) coefficient  $\alpha$ , except that it relaxes the assumption of essential tau equivalence (i.e., an assumption that all items have equal factor loadings onto the latent construct).<sup>1</sup> Reliability for the EF factors was weak but acceptable and increased across the four waves of the present study ( $\omega$  = .69, .74, .74, .78, for Waves 1 through 4, respectively).

**Measures of academic achievement.** Children's literacy and math skills were assessed using the Woodcock Johnson Psycho-Educational Battery-III Tests of Achievement (WJ-III; Woodcock, McGrew, & Mather, 2001a) in English or the Bateria III Woodcock-Muñoz (Muñoz-Sandoval, Woodcock, McGrew, & Mather, 2005a) in Spanish. In a study using a large and diverse sample of 2000 children, cross-language equating procedures were employed using item-response theory (IRT) methods. Results suggested that the WJ-III and the Woodcock-Muñoz assess the same competencies and can be combined appropriately for use in cross-language studies (Woodcock & Muñoz-Sandoval, 1993). Woodcock-Johnson W-scores were used because they utilize Rasch-based measurement models to create equal-interval scale characteristics, with the W-score centered at 500 as the approximate average performance of a 10-year-old (Mather & Woodcock, 2001).

**Letter-Word Identification.** Children's literacy skills were measured using the Letter-Word Identification subtest of the WJ-III (Woodcock et al., 2001a) or the Bateria III Woodcock-Muñoz (Muñoz-Sandoval et al., 2005a). This test measures letter identification and word-reading skills through expressive and receptive items and had strong internal consistency for both the English-speaking (Cronbach's alpha: Wave 1 = .92, Wave 2 = .92, Wave 3 = .94, Wave 4 = .94) and Spanish-speaking children (Cronbach's alpha: Wave 1 = .83, Wave 2 = .80, Wave 3 = .83, Wave 4 = .90) in the present sample. Although these two subtests have been deemed comparable in rigorous cross-language validation studies in terms of content and difficulty (Woodcock et al., 2001b), they could not be appropriately combined to provide full-sample

reliabilities because children receive different items to ensure cultural relevance.

**Applied problems.** Children's math skills were measured using the Applied Problems subtest of the WJ-III (Woodcock et al., 2001a) or the Bateria III Woodcock-Muñoz (Muñoz-Sandoval et al., 2005a). This measure assesses children's early mathematical operations (e.g., counting, addition, and subtraction) through practical problems and had good internal consistency for the full sample (Cronbach's alpha: Wave 1 = .86, Wave 2 = .87, Wave 3 = .85, Wave 4 = .83), for English-speaking children only (Cronbach's alpha: Wave 1 = .80, Wave 2 = .81, Wave 3 = .79, Wave 4 = .81), and for Spanish-speaking children only (Cronbach's alpha: Wave 1 = .86, Wave 2 = .82, Wave 3 = .82, Wave 4 = .80).

## Analytic Approach

We examined longitudinal relations between EF and two academic domains: math and literacy. As described above, we explored these relations using two separate sets of analyses (i.e., cross-lagged panel models and LGCM) to obtain a more complete understanding of our data than could be provided by either analysis alone. Although we had specific hypotheses for our research questions, we did not favor one analytic approach over the other when interpreting the models that were used to answer our research questions. Instead, we chose two analytic models because each model provides a unique perspective on the data at hand and to our overarching research question. Treating each model as equally informative allows for a fuller understanding of the developmental processes impacting EF and academic achievement.

Participating children were nested in classrooms at each wave, and we computed ICCs for all target variables (i.e., EF indicators as well as math and literacy scores). The models for these ICCs specified wave-specific clustering, such that ICCs for Wave 1 variables used Wave 1 classrooms as the clustering units, ICCs for the Wave 2 variables used Wave 2 classrooms as the clustering units, et cetera. We anticipated that between-classroom differences would be strongly related to sociodemographic factors, and we supplemented our examination of ICCs with computation of conditional ICCs. To obtain conditional ICCs, we first regressed all EF and academic achievement variables on participant age (at Time 1), Head Start Status, and ELL status in a single-level regression model and stored the residuals from these models (i.e., residual centering). We then fit a saturated two-level path analysis (i.e., freely estimating all item variances and covariances at both levels) for each wave of data and obtained conditional ICCs for the EF and academic achievement variables.

As Table 1 shows, all variables exhibited substantial variability at the between-classroom level (i.e., all ICCs > .05), but this variance was largely accounted for by the demographic covariates. Only kindergarten literacy retained a substantial amount of between-classroom variance. Appropriately modeling the longitu-

<sup>1</sup> Willoughby, Pek, and Blair (2013), have advocated for the use of maximal reliability—the reliability of an optimally weighted composite—when examining latent EF factors. However, recent simulation evidence has drawn the usefulness of maximal reliability into question (Geldhof, Preacher, & Zyphur, 2014). We therefore do not include estimates of maximal reliability in the present study.



Table 1  
ICCs for All Items by Wave-Specific Cluster

Construct indicator	Unconditional				Conditional			
	W1	W2	W3	W4	W1	W2	W3	W4
Executive function								
Working memory	.09	.12	.14	.12	.04	.00	.01	.06
Simon says	.06	.09	.10	.11	.02	.02	.03	.06
HTKS	.12	.19	.13	.10	.06	.03	.02	.02
Card sort	.11	.11	.13	.07	.03	.00	.01	.02
Literacy	.21	.21	.17	.12	.05	.02	.14	.15
Math	.19	.17	.19	.19	.01	.04	.05	.03

Note. Conditional intraclass correlations (ICCs) control for age (at Time 1), Head Start status, and English language learner (ELL) status. HTKS = Head-Toes-Knees-Shoulders task.

dinal observations as nested in children and also cross-classified in wave-specific classrooms would complicate our results and detract from model interpretability. Given that controlling for the covariates using single-level regression largely accounted for between-classroom variation in the measures, we present single-level models that control for the same covariates included when computing conditional ICCs. The caveat, therefore, is that the standard errors for paths involving kindergarten literacy may be slightly biased.

**Data screening.** Both sets of analyses used robust maximum likelihood estimation (MLR in Mplus) to deal with non-normality and missing data (Muthén & Muthén, 1998–2015). Skewness ranges for the four EF tasks and achievement measures were –1.23 to 1.97 at Wave 1, –1.20 to 0.83 at Wave 2, –1.85 to 1.02 at Wave 3, and –2.37 to 0.74 at Wave 4. Kurtosis ranges were 1.29 to 5.87 at Wave 1, 1.72 to 6.16 at Wave 2, 1.44 to 7.07 at Wave 3, and 1.63 to 8.58 at Wave 4. For children participating in the study at any given wave (i.e., missing data not due to children leaving the study in between waves of data collection), there was less than 6% missing data on direct assessments and no missing data on age, gender, Head Start status, or language status (see Table 2 for the number of observations for every variable). Once missing data due to attrition was factored in (i.e., children leaving the longitudinal study and resulting in a loss of data at later waves), the range of missing data was 0–30.66% for individual measures (average missingness was 15.34%). Two variables had 30.66% missing data at wave four (i.e., the Simon Says task and Auditory Working Memory; 294 observations out of the original sample size of 424). Because most missing data occurred due to participant attrition, we created binary variables (0 = did not leave study, 1 = did leave study) to test whether any of our covariates were related to attrition throughout the study. None of our covariates were related to attrition that occurred within the school years (i.e., Wave 1 to Wave 2 and Wave 3 to Wave 4). For attrition between Waves 2 and 3 (i.e., the transition from prekindergarten to kindergarten), we found that Head Start status ( $b = 0.72, p = .005$ ) and parent education ( $b = -0.08, p = .015$ ) were significantly related to attrition when running bivariate logistic regression models. In other words, children in Head Start and children of parents with fewer years of education were more likely to leave the study between the spring of prekindergarten and fall of kindergarten. However, when both predictors were used to predict attrition, neither was significant, suggesting substantial shared variance in their relation to attrition (i.e., reduction in size of coefficient and

increases in standard errors). Thus, all models included Head Start status (as opposed to parent education, which had substantial missing data), along with child age and language status as time-invariant covariates.

**Cross-lagged panel model.** We used a cross-lagged panel model to examine whether changes in relative standing on each construct (EF, math, literacy) were related over time. The cross-lagged panel models specifically tested whether children whose EF, math, and literacy scores were higher (or lower) than their peers at earlier waves were also higher (or lower) than their peers at subsequent times of measurement (i.e., a test of stability). After controlling for the stability of relative standing, these analyses also allowed us to test whether relative standing on one variable at earlier waves predicted *changes in relative standing* (not changes in absolute magnitude) on a different variable at subsequent waves. For each cross-lagged effect (e.g., EF predicting changes in math), we simultaneously examined the reciprocal relation (e.g., math predicting changes in EF) as a test of bidirectionality.

We first specified a longitudinal confirmatory factor analysis (CFA), controlling all indicators for participants’ age at the beginning of the study, ELL status, and Head Start enrollment status. This approach to controlling for covariates allows minor differences between indicators and the covariates to not impact overall model fit (see also Geldhof, Pornprasertmanit, Schoemann, & Little, 2013). The initial CFA allowed us to examine the structure of EF because, although there is strong evidence in younger children that EF is best described as a unitary construct (Hughes, Ensor, Wilson, & Graham, 2009; Wiebe, Espy, & Charak, 2008), there is also evidence that it becomes more differentiated over time (Huizinga, Dolan, & van der Molen, 2006; Lehto, Juujärvi, Kooistra, & Pulkkinen, 2003). Good fit for a CFA that specified a single EF factor per time point would support the underlying assumption of our analyses—that EF is reasonably unidimensional as it was measured in this sample.

We scaled all latent variables in the initial CFA by fixing latent means of zero and latent variances to one. We modeled math and literacy as single-indicator factors by freely estimating the factor loading for each indicator onto its respective construct and additionally fixing the indicators’ residual variances to zero. To account for correlated residuals over time, we estimated residual covariances within each indicator of EF (e.g., all HTKS indicators were allowed to covary, independent of their relations implied by the stability of EF as a latent construct). Figure B1 in the Appendix provides a partial path diagram that illustrates the EF component of this model.

We established measurement invariance of the EF construct across waves using the change in confirmatory fit index (CFI) criterion suggested by Cheung and Rensvold (CFI decreases by  $< .01$ ; 2002). Modeling invariance requires equating factor loadings (weak invariance) and intercepts (strong invariance), allowing for differences at the latent level (i.e., latent variances and means, respectively; see Little, 1997 for a discussion). Thus, latent variances for EF in Times 2 through 4 were freely estimated in the weak invariance model, and latent means for EF in Times 2 through 4 were additionally freed in the strong invariance model. These tests ensured that the qualitative meaning of EF remained stable across the four waves of data collection rather than EF being strongly indicated by one measure in earlier waves and strongly indicated by a different measure in later waves. Invariance could



Table 2  
Descriptive Statistics for All Study Variables

Variable	Prekindergarten (Year 1)				Kindergarten (Year 2)			
	Fall (Wave 1)		Spring (Wave 2)		Fall (Wave 3)		Spring (Wave 4)	
	<i>N</i>	<i>M</i> ( <i>SD</i> )	<i>N</i>	<i>M</i> ( <i>SD</i> )	<i>N</i>	<i>M</i> ( <i>SD</i> )	<i>N</i>	<i>M</i> ( <i>SD</i> )
Age	424	4.70 (0.30)	394	5.15 (0.30)	308	5.67 (0.30)	299	6.17 (0.29)
Percent male	424	51%	394	51%	308	50%	299	51%
Percent Head Start	424	55%	394	54%	308	51%	299	51%
Percent ELL	424	15%	394	15%	308	15%	299	14%
Parent education	353	14.40 (4.23)	336	14.34 (4.26)	275	14.60 (4.45)	269	14.67 (4.46)
HTKS	403	17.41 (17.20)	391	25.15 (18.28)	303	33.17 (17.74)	296	39.22 (16.00)
Card sort	409	13.64 (6.67)	389	16.49 (5.92)	307	18.60 (4.88)	295	19.78 (3.88)
Working memory	400	450.30 (14.80)	385	456.17 (17.97)	303	464.60 (19.21)	294	473.18 (19.90)
Simon Says	408	0.14 (0.28)	387	0.29 (0.38)	307	0.45 (0.39)	294	0.54 (0.38)
Math	401	410.17 (23.30)	391	419.83 (23.11)	305	431.02 (20.71)	295	442.09 (19.29)
Literacy	408	335.65 (26.59)	390	349.33 (26.80)	305	366.00 (29.14)	295	400.24 (35.21)

Note. ELL = English language learner; HTKS = Head-Toes-Knees-Shoulders task; Working memory = Auditory Working Memory subtest from the Woodcock-Johnson III Tests of Cognitive Abilities; Math = Applied Problems subtest from the Woodcock-Johnson III Tests of Achievement; Literacy = Letter-Word Identification subtest from the Woodcock-Johnson III Tests of Achievement.

not be tested for math or literacy because those factors had only one indicator per time point (e.g., equating factor loadings for math over time would result in three additional degrees of freedom that would then be lost by freely estimating the latent variances for math at Times 2 through 4, resulting in no change in model fit).

After establishing measurement invariance, we specified a longitudinal structural equation model (SEM) that included single-lag stability regressions (e.g., EF at Time 1 predicting EF at Time 2) and single-lag cross-construct regressions (e.g., EF at Time 1 predicting math at Time 2). We freely estimated all within-wave covariances (e.g., EF at Time 1 covaried with math and literacy at Time 1). The cross-lagged panel model assumes no longitudinal covariances except those specified by the longitudinal regression coefficients, and we tested this assumption by first estimating all covariances between constructs separated by more than one lag (e.g., EF at Time 1 covaried with math at Times 3 and 4). The latent variable covariance matrix was therefore completely saturated, and our initial SEM model had identical fit to our strong-invariance CFA model. We then tested the assumption of no longitudinal covariance by removing all covariances between constructs separated by more than one lag and performing a likelihood ratio test.

**LGCM.** To examine whether rates of change in EF, math, and literacy were correlated in our data, we estimated the associations between the growth parameters for each construct in a three-trajectory LGCM. Based on the assumption that growth in the target variables, especially EF (Zelazo et al., 2013), may be non-linear, we specified quadratic growth curves for all target constructs. The model then examined how initial standing (i.e., the random intercepts) and the rates of change and acceleration (i.e., the random linear and quadratic slopes) were correlated. The LGCM treated EF as a latent factor, meaning the growth model for that construct was technically a curve-of-factors model (McArdle, 1988; see also Hancock, Kuo, & Lawrence, 2001). We imposed the same invariance constraints from the panel model on the EF factor in the growth model, although factors in the growth model were identified by constraining the factor loading of HTKS to one and fixing the intercepts for all HTKS indicators to zero. Latent

intercepts for EF were also fixed to zero to identify the growth component of the model (see also Figure 1 in Hancock et al., 2001). Due to model complexity, and to acknowledge that the covariates were between-persons variables, we controlled for all covariates at the level of the growth parameters. Figure B2 in the Appendix provides a partial path diagram of the EF component of this model.

For the sake of comparability to our panel models, we used wave in the study as loadings for these models (i.e., loadings for the linear slope were 0, 1, 2, and 3, for Waves 1, 2, 3, and 4, respectively). This approach allowed us to model each wave of data as a discrete time point rather than taking the more traditional approach of modeling each observation of each child as occurring at the child's unique age at the assessment. Given participants' relatively narrow age range, very few children in later waves were younger than children measured in earlier waves. That is, child ages did not substantially overlap across waves.

## Results

Descriptive statistics are presented in Table 2. Overall, and as expected, children improved at each wave of the study on EF tasks, math, and literacy.

## Panel Models

The initial CFA fit the data well (fit for all panel models is presented in Table 3) and had statistically significant factor loadings for all indicators of EF (all  $ps < .001$ ). Modification indices did not indicate areas of localized misfit. An initial test of weak (i.e., loading) invariance substantially decreased model fit ( $\Delta CFI = -.02$ ), with modification indices suggesting that the relation between EF and the Card Sort total score changed across waves and that the relation between Working Memory and EF was significantly different at Wave 4 than in the other waves. Freely estimating the Card Sort factor loading for waves 1 and 2 and the Working Memory factor loading in Wave 4 resulted in a model that supported partial weak invariance ( $\Delta CFI = -.005$ ;  $\Delta$  Bayesian

Table 3  
*Fit for Panel Models*

Panel models	$\chi^2$ <sup>a</sup>	df	RMSEA	90% CI (RMSEA)	CFI	TLI	BIC
Initial	284.72	170	.04	[.03, .05]	.979	.96	30786.98
Weak invariance	387.09	179	.05	[.05, .06]	.962	.93	30831.28
Partial weak invariance	319.45	176	.04	[.04, .05]	.974	.95	30783.33
Strong invariance and initial SEM <sup>b</sup>	343.63	185	.05	[.04, .05]	.971	.95	30754.58

*Note.* RMSEA = root-mean-square error of approximation; CI = confidence interval; CFI = comparative fit index; TLI = Tucker–Lewis Index; SEM = structural equation model; BIC = Bayesian information criterion.

<sup>a</sup> Models were estimated using robust maximum likelihood; chi-squared statistics cannot be directly compared. <sup>b</sup> Fit for these models was identical because the latent covariance structure of the initial SEM was saturated.

information criterion [BIC] =  $-3.65$ ). Equating the intercepts across waves in this model further supported partial strong factorial invariance ( $\Delta$  CFI =  $-.003$ ;  $\Delta$  BIC =  $-28.75$ ). Table 4 presents factor loadings and intercepts from the strong invariance model and highlights which parameters were freely estimated versus equated across time. This model shows that HTKS was a relatively strong indicator of EF across waves, Working Memory was a stronger indicator of EF in Wave 4 (relative to other waves), and the Card Sort was an especially strong indicator of EF at Wave 1. Latent variances and correlations from this model are presented in Table 5, with actual (rather than latent) means and variances provided for the math and literacy scores.

Table 4  
*Factor Loadings and Intercepts from the Strong Invariance CFA Model*

Construct indicator	Standardized loading (SE) <sup>a</sup>	Raw-metric loading (SE) <sup>a</sup>	Raw-metric intercepts (SE)
Time 1 EF			
Working memory	.34* (.04)	.51 (.06)	45.37 (.10)
Simon Says	.45* (.04)	.67 (.07)	.90 (.10)
HTKS	.45* (.06)	.49 (.06)	1.36 (.08)
Card sort	.59* (.04)	1.30 (.09)	5.04 (.15)
Time 2 EF			
Working memory	.40* (.03)	equated (T1)	equated (T1)
Simon Says	.49* (.03)	equated (T1)	equated (T1)
HTKS	.58* (.04)	equated (T1)	equated (T1)
Card sort	.55* (.05)	.76 (.09)	equated (T1)
Time 3 EF			
Working memory	.41* (.03)	equated (T1)	equated (T1)
Simon Says	.52* (.04)	equated (T1)	equated (T1)
HTKS	.63* (.04)	equated (T1)	equated (T1)
Card sort	.52* (.05)	.54 (.07)	equated (T1)
Time 4 EF			
Working memory	.54* (.04)	.72 (.09)	equated (T1)
Simon Says	.54* (.04)	equated (T1)	equated (T1)
HTKS	.66* (.04)	equated (T1)	equated (T1)
Card sort	.59* (.04)	equated (T3)	equated (T1)

*Note.* Indicators were divided by constants to make their variances more homogenous, expediting model convergence (e.g., Muthén, 2010). For more information, see Appendix A. HTKS = Head–Toes–Knees–Shoulders task; EF = executive function; CFA = confirmatory factor analysis.

<sup>a</sup> Loadings are somewhat attenuated because covariates were controlled at the item level.

\*  $p < .001$ .

We next specified a cross-lagged panel SEM and tested the assumption of no longitudinal covariances above and beyond those specified by the lag-1 structural regressions. A likelihood ratio test comparing models that did versus did not allow longitudinal covariances between EF, math, and literacy measures separated by two or more lags (e.g., EF at Time 1 and math at Time 3) supported this assumption,  $\Delta \chi^2(27) = 30.70$ ,  $p = .28$ ;  $\Delta$  BIC =  $-130.28$ . The structural component of this final model is illustrated in Figure 1. This path diagram omits nonsignificant regression estimates and within-wave covariances. These additional details are provided in Figure B3 and Table B1 of the Appendix.

Results suggest that (a) relative standing on all variables was stable (i.e., all autoregressive paths were statistically significant at  $p < .001$ ), with EF displaying especially high stability ( $\beta$ s ranged from .75 to .86); (b) that changes in relative standing on EF and literacy were essentially unrelated across waves (i.e., low cross-lagged regression coefficients); (c) that EF and math were mutually influential in preschool and this relation shifted in kindergarten, such that only EF predicted math; and (d) that math and literacy were not consistently related across time.

## LGCM

The initial three-trajectory model indicated a non-positive-definite latent covariance matrix caused by collinearity between the EF intercept and quadratic slope and by nonsignificant residual variances for the linear slope for literacy and the quadratic slope for math. These nonsignificant residual variances suggest an overfitted model. We therefore eliminated the collinearity by regressing the quadratic slope for EF on the intercept for EF and constraining the residual variance of the quadratic slope to zero. We also fixed the nonsignificant residual variances to zero. These constraints did not significantly reduce model fit,  $\Delta \chi^2(23) = 29.68$ ,  $p = .16$ ;  $\Delta$  BIC =  $-107.65$ , and the resulting model fit the data well,  $\chi^2(277) = 574.57$ ,  $p < .001$ ; RMSEA = .05 [.05, .06]; CFI = .95; TLI = .93. An examination of the modification indices did not reveal areas of extreme local misfit. Table 6 contains the estimated means and variances for the latent growth parameters from this model and clarifies which growth parameters were estimated in which ways (i.e., fixed vs. random variances). All growth parameters with freely estimated variances were regressed on the covariates and allowed to covary among themselves. Partial correlations among these parameters are presented in Table 7. Fixed growth parameters were regressed on the control variables (i.e., age, Head Start status, and ELL status) but did not have freely



Table 5  
Means, Variances, and Correlations for Strong Invariance Model

Constructs	1	2	3	4	5	6	7	8	9	10	11	12	M
1. EF1	—												.00
2. Math1	.65*	4.03											41.71
3. Literacy1	.44*	.39*	5.45										34.77
4. EF2	.91*	.71*	.43*	1.97									1.50
5. Math2	.68*	.74*	.40*	.74*	3.72								42.71
6. Literacy2	.36*	.41*	.77*	.40*	.41*	5.66							36.16
7. EF3	.85*	.70*	.40*	.92*	.78*	.34*	2.36						2.63
8. Math3	.59*	.67*	.41*	.66*	.79*	.40*	.77*	3.14					43.67
9. Literacy3	.36*	.37*	.70*	.40*	.35*	.80*	.36*	.41*	7.02				37.60
10. EF4	.71*	.64*	.32*	.83*	.69*	.32*	.92*	.71*	.34*	2.18			3.24
11. Math4	.60*	.62*	.41*	.67*	.70*	.41*	.71*	.75*	.41*	.69*	2.79		44.74
12. Literacy4	.43*	.42*	.60*	.46*	.45*	.65*	.43*	.50*	.78*	.43*	.48*	10.69	41.02

Note. Variances on diagonal, correlations below diagonal. Actual (as opposed to latent) means and variances provided for Math and Literacy. Indicators were divided by constants to make their variances more homogenous, thus expediting model convergence (e.g., Muthén, 2010). For more information, see Appendix A. EF = executive function.

\* $p < .001$ .

estimated variances and did not covary with any other growth parameter. The correlations in Table 7 highlight strong associations among the intercepts and between the intercept and slope parameters. Average growth trajectories are plotted in Figure 2.

To better understand how the constructs at Wave 1 (estimated as the intercepts) may have impacted the results of the LGCMs, we took the additional step of regressing all three random slopes on all three random intercepts. This final model allowed us to examine how absolute changes in each variable were correlated after controlling for initial standing on each (i.e., all three random intercepts). As shown in Table 8, the residual random slope for math was significantly correlated with both other slopes, although the correlation between the EF and literacy slopes was not statistically significant. Thus, after children's initial standing was accounted for in the LGCM, the results suggested growth in EF and math were associated during this developmental period.

## Discussion

The overarching aim of the current study was to examine the longitudinal relations between EF, math, and literacy across four waves of measurement spanning preschool and kindergarten. We employed a multi-analytic approach, first using a cross-lagged panel model to test the extent to which relative standing on EF, math, and literacy were related across time. We then used LGCMs to test whether growth in our constructs were associated. As expected, results generally demonstrated significant reciprocal relations and correlated growth between EF and math as well as math and literacy, but not between EF and literacy. Notably, results from our panel models indicated that these significant relations may change over time. For example, EF predicted math but math did not predict EF during the kindergarten school year. These findings contribute to the current literature by demonstrating

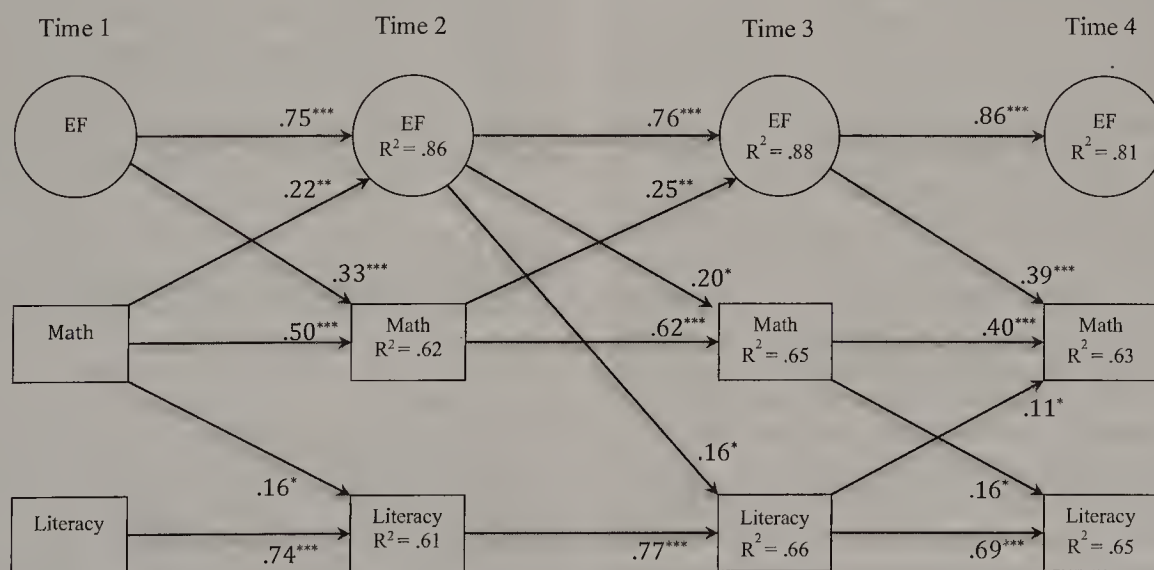


Figure 1. Path diagram for our final structural model (standardized coefficients). Within-wave covariances and nonsignificant regression paths not shown). Time 1 = fall of preschool; Time 2 = spring of preschool; Time 3 = fall of kindergarten; Time 4 = spring of kindergarten.

Table 6  
Estimated Growth Parameter Conditional Means and Variances

Parameter	M (SE)	Variance (SE)
Executive function <sup>a</sup>		
Intercept	1.299 (.07)***	.28 (.05)***
Linear	.91 (.07)***	.05 (.02)*
Quadratic	-.13 (.02)***	.00 (FIXED)
Literacy		
Intercept	34.86 (.18)***	4.26 (.49)***
Linear	.60 (.17)***	.00 (FIXED)
Quadratic	.48 (.05)***	.05 (.01)***
Math		
Intercept	41.72 (.14)***	3.18 (.51)***
Linear	.97 (.12)***	.10 (.03)**
Quadratic	.01 (.04)	.00 (FIXED)

Note. Indicators were divided by constant values to create more homogenous indicator variances. Values in this table therefore provide meaningful information about the shape of each growth trajectory but do not describe scores in their raw metric.

<sup>a</sup> Calculated as the estimated intercept (.02) plus the conditional mean of the executive function (EF) intercept (1.299) multiplied by the regression coefficient regressing the EF quadratic slope in the EF intercept (-.12).

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

a bidirectional association and correlated growth between EF, a more domain-general set of cognitive processes, and math, a domain-specific skill. These results have implications for research on curriculum development and intervention design. Further, this study adds to the theoretical discourse surrounding the development of EF and academic skills in early childhood.

### Bidirectional Relations Between EF, Math, and Literacy: Cross-Lagged Panel Models

Consistent with previous research (Blair & Razza, 2007; Bull et al., 2008; Bull, Johnston, & Roy, 1999; McClelland et al., 2007), our panel models suggested that EF is a significant predictor of math in preschool and kindergarten. These findings provide support for the notion that EF may be foundational for the development of important early math skills. In addition, and also consistent with a recent study (Fuhs et al., 2014), these results demonstrated reciprocal associations between EF and math during preschool and as children transition into kindergarten (i.e., from the spring of preschool to the fall of kindergarten). These findings suggest that EF may not only be important for the development of math, but that math may also be important for the development of EF during this time. Thus, it is possible that math skills are foundational for growth in EF. Essentially, the ability to pay attention, remember complex rules, and persist on challenging tasks may help children perform better on math tasks (McClelland et al., 2007) and, conversely, strong math skills (e.g., solving complicated math problems) may contribute to children's ability to sustain attention, remember a series of rules, and inhibit incorrect responses on complex EF tasks (Fuhs et al., 2014).

With the addition of a fourth time point at the beginning of kindergarten (in comparison to prior research), we were able to extend the existing literature and identify at which point during preschool and kindergarten relations between EF and math may change. Findings revealed that although math and EF were reciprocally related during preschool and during the transition to kin-

dergarten, this bidirectional relation faded during the kindergarten year. Specifically, during the kindergarten year (between Waves 3 and 4), EF in the fall remained a significant predictor of math in the spring, but not vice versa. Changes in the relations between EF and math may be due to factors associated with preschool and/or kindergarten instruction. In kindergarten, children are charged with more challenging math tasks and they may need to call upon EF skills to resist the natural inclination to either give up and abandon a task or use a less efficient previously learned rule (Bull et al., 1999). In contrast, mathematics instruction in preschool is often limited in complexity and focused around a narrow range of activities (e.g., counting; Ginsburg, Lee, & Boyd, 2008). It may be the case that, in preschool—where limited mathematics instruction is provided—children who have higher levels of math skills are engaged in instructional activities that provide the opportunity for them to develop higher EF skills and, in turn, those children with higher levels of EF are better able to acquire the limited mathematics instructional information that is provided. These instructional differences may explain why the bidirectional relationship (EF  $\longleftrightarrow$  math) emerges during preschool and fades during kindergarten, when children begin experiencing more uniform and frequent math instruction during kindergarten.

Taken together, these findings have potential implications for the development and evaluation of instructional strategies and interventions that are designed to improve either EF or math. In preschool, it may be more beneficial for children if teachers target both EF and math simultaneously, whereas in kindergarten, focusing instructional efforts on EF as a foundational skill set may be more important. Additionally, these findings suggest that children who enter kindergarten with low levels of EF may be at risk for academic difficulties and in need of extra instructional supports or intervention. Critically, the causal nature of such instructional strategies need to be evaluated experimentally.

In contrast to math, the panel model indicated that relative standing on EF and literacy were essentially unrelated across waves. These findings are not surprising given inconsistent links between EF and literacy in previous studies (Blair & Razza, 2007; Blair et al., 2015; Cameron Ponitz et al., 2009; Schmitt et al., 2014) and nonsignificant bidirectional associations in recent work (Fuhs et al., 2014). Several speculations as to why associations are stronger for EF and math than EF and literacy have been introduced in recent literature. For example, some argue that math content and activity place more cognitive demands on children and, thus, require stronger EF skills to master (Bull et al., 2008;

Table 7  
Partial Correlations Between Growth Parameters

Parameters	1	2	3	4	5	6
1. EF—Intercept	—					
2. EF—Linear	.82***	—				
3. Literacy—Intercept	.48***	.38**	—			
4. Literacy—Quadratic	.20**	.31**	.08	—		
5. Math—Intercept	.81***	.82***	.49***	.25*	—	
6. Math—Linear	-.40**	-.17	-.14	.12	-.53***	—

Note. The raw-metric regression of the executive function (EF) quadratic slope on the EF intercept was  $-.12$  ( $p < .01$ ). A parallel model that relied on numerical integration provided a standardized coefficient of  $-.91$ .

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



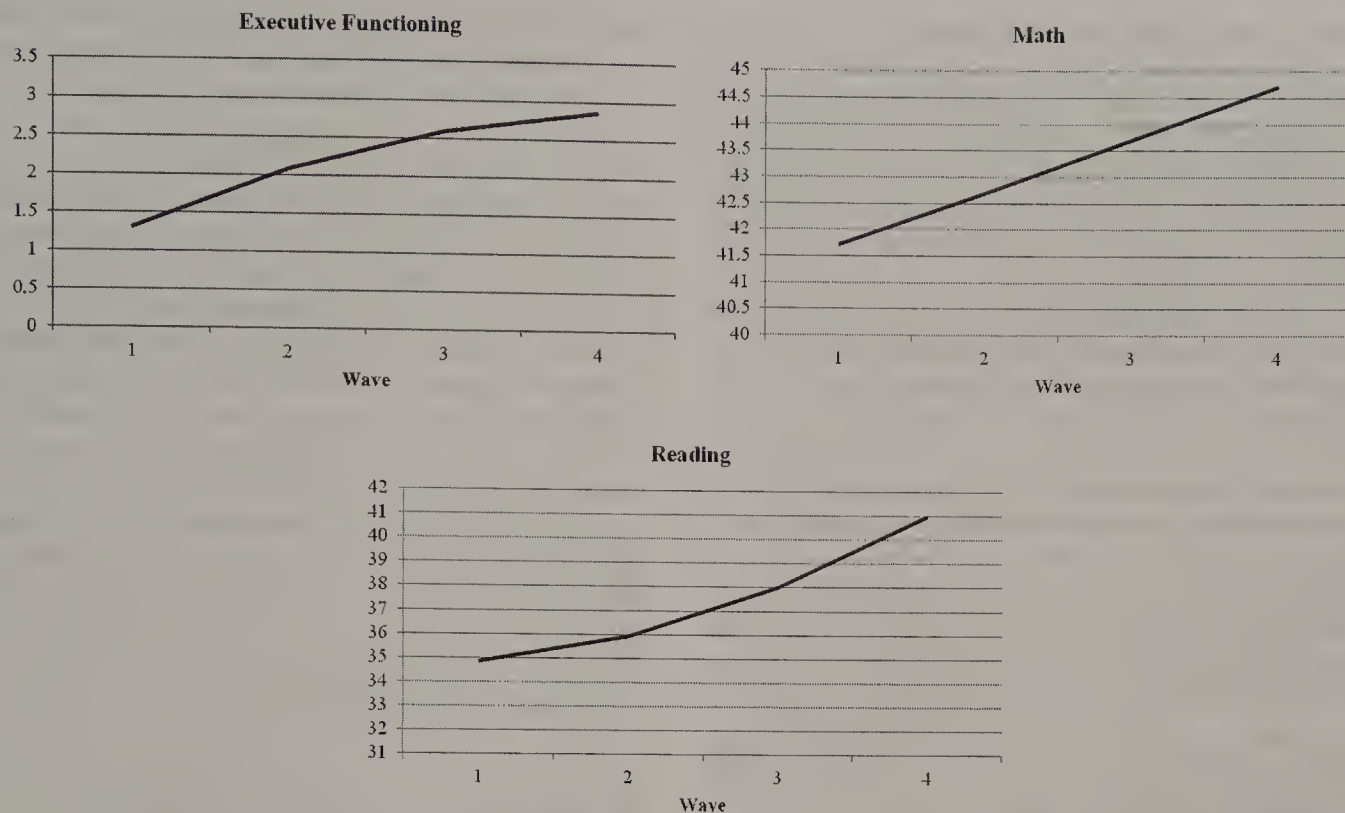


Figure 2. Average growth trajectories from the latent growth curve modeling (LGCM). Math and literacy scores were rescaled from original values.

Clark et al., 2010; Espy et al., 2004; Willoughby, Blair, et al., 2012). A second explanation is that EF is a foundational skill set that supports growth in reasoning abilities (Richland & Burchinal, 2013). Higher-order reasoning skills are necessary to succeed on math tasks that require children to solve complex story or word problems (e.g., “Katie had three balls. One of them rolled away. Now how many does she have?”; Blair et al., 2015). In contrast, literacy tasks typically assess children’s knowledge, making fewer demands on reasoning abilities and EF. Others argue that differences in academic focus in early childhood classrooms could play a role in explaining differences in the development of math versus literacy (Cameron Ponitz et al., 2009). Extant research suggests that preschool teachers spend more time engaged in literacy instruction compared to math instruction (Layzer, Goodsen, & Moss, 1993; Skibbe, Hindman, Connor, Housey, & Morrison, 2013). Children may, therefore, have to engage in math activities (e.g., patterning during free play) spontaneously and independently during the school day, which may require higher levels of EF. Similarly, parents report engaging in significantly more literacy activities at home than math activities (Cannon & Ginsburg, 2008; Skwarchuk, Sowinski, & LeFevre, 2014). Parents who believe

their children are more academically ready may engage their children in more cognitively demanding math activities at home (DeFlorio & Beliakoff, 2015). The greater consistency of literacy activities at home and school may contribute to its overall distinction from growth in EF.

Another aspect of our research question was to investigate bidirectional relations between math and literacy skills across the preschool and kindergarten years. Somewhat contrary to our expectations, these relations were weaker in preschool and became bidirectional during the kindergarten year. These differences in the findings compared to expectations also are likely due to instructional practices. In contrast to the divergence of the relation between math and EF, there may be a convergence in the relation between math and literacy as instruction in both domains becomes more parallel in quantity. In preschool, children are generally exposed to more literacy instruction compared to math instruction. In contrast, in kindergarten, math and literacy instruction become more uniform and consistent, and all children are typically exposed to the same quantity of instruction for both academic domains. This parallel exposure likely allows children to draw on concepts learned from the instruction in the other domain (e.g., being able to read a word problem allows children to complete the math task) and thus, the relation between math and literacy may be strengthened.

### Correlated Growth Between EF, Math, and Literacy: LGCMs

To further investigate the longitudinal associations between EF, math, and literacy, we employed a second analytic approach: LGCM. Consistent with prior evidence (e.g., McClelland et al., 2007), these models indicated that the latent intercepts (a proxy for

Table 8  
Correlations Between Slopes, Conditional on Intercepts  
and Covariates

Parameter	1	2	3
1. EF—Linear	—		
2. Literacy—Quadratic	.21	—	
3. Math—Linear	.63**	.32**	—

\*\*  $p < .01$ .

where children started) for EF, math, and literacy were all significantly correlated, suggesting that performance relative to peers was consistent across measures. Also, consistent with prior research (McClelland et al., 2007; Schmitt et al., 2014), initial levels of EF and math were more highly correlated than EF and literacy. However, whether the coupling of the three variables is a result of unidirectional causality, bidirectional causality, or the result of unmeasured third variables is not clear.

In terms of cross-domain relations in growth, the final LGCM indicated that, after controlling for initial standing on each construct (i.e., all latent intercepts), the latent EF and math slopes were positively correlated. In contrast, results revealed a nonsignificant relation between growth in EF and literacy. This finding is in line with prior studies demonstrating that the longitudinal association between EF and math is more robust than EF and literacy (Blair et al., 2015; Cameron Ponitz et al., 2009). This finding also supports our earlier assertion that engaging in math activities may be a context in which children are able to expand their EF and that domain-specific differences in instruction during the preschool and kindergarten years may account for these differential patterns of growth. For example, over the last two decades, there has been a strong emphasis on early literacy instruction in both preschool and kindergarten. Indeed, previous research indicates a strong schooling effect for children's literacy development (Burrage et al., 2008; Christian, Bachman, & Morrison, 2001). Due to this emphasis on literacy instruction, children may not need to call upon their EF as much when engaging in literacy activities, and thus, improvement in EF would be less likely to be related to improvement in literacy during this time frame. Finally, the latent math and literacy slopes were significantly related, providing additional evidence that early math and literacy skills codevelop over the preschool and kindergarten years (Duncan et al., 2007; LeFevre et al., 2010; Purpura et al., 2011).

### Conclusions From the Integration of Both Analytic Approaches

Results from the two analytic approaches provide a similar story with regard to our overarching research question. Both the panel model and the LGCM suggested positive correlations between initial levels of EF, math, and literacy. Thus, and consistent with previous research (McClelland et al., 2007; Schmitt et al., 2014), there is strong evidence that these three constructs are tightly coupled by the time children enter preschool. However, both sets of results also suggest that EF and math are consistently related over time, whereas the association between EF and literacy is weak. Taken together, the LGCM and panel model therefore suggest that some early factor (math or an outside variable) likely helps explain the correlation between EF and literacy. The development of EF and literacy seem to be driven by separate processes during the transition to kindergarten, however.

### Limitations and Future Directions

Although this study extends existing literature on the relations between EF and early academic skills, there are also several limitations. First, we utilized several measures of EF in our study but only one measure each for math (Applied Problems) and literacy (Letter-Word Identification). These subtests measure spe-

cific components of math (e.g., counting, calculation) and literacy (e.g., decoding, word-reading) and may therefore not represent comprehensive growth in these broader academic domains. It will be important for future studies to include additional measures of early academic skills to further our understanding of how complex skills like math and literacy develop. For instance, other research has shown that the relations between EF and math differ based on the distinct subcomponents of math that were measured (Lan, Legare, Cameron Ponitz, & Morrison, 2011; Purpura & Ganley, 2014). A comparison of more targeted relations was not possible in the current study due to our use of only one measure each for math and literacy. Moreover, utilizing multiple measures of math in future studies will help elucidate the extent to which EF actually differentially predicts components of math at different ages. Indeed, as the Applied Problems subtest becomes more challenging, demands on EF become stronger. Changes in the relations between EF and math at different ages may not necessarily mean EF is a better or worse predictor of math, but that changes in these relations are related to the mathematics concepts targeted within specific assessment measures.

Second, as noted above, the quantity of instruction may have varied across time for specific domains (particularly for math), and these differences may have altered the relations between domains. For example, more time spent engaging in math instruction may affect the development of math, which, in turn, could change the relations between math and EF or between math and literacy. In the current study, math and literacy instructional practices, activities in schools and at home, or active learning in these domains were not assessed. Moreover, other contextual factors as well as individual child characteristics not measured in this study, such as parenting practices, early language abilities, or motor development, may be contributing to growth in EF and academic skills (McClelland et al., 2015). Further research that includes contextual factors and additional child characteristics may enhance our understanding of the linked development across these domains.

Third, recent research suggests that cross-lagged panel models can produce biased estimates due to unmodeled trait-like stability (e.g., Hamaker, Kuiper, & Grasman, 2015). Although the present analyses used a likelihood ratio test to show no evidence of additional trait-like stability (i.e., by constraining the correlations between factors separated by more than one lag to be zero), it will be critical for future research to explore alternative model specifications when investigating EF and academic outcomes over time. Future studies should also test for mediating effects (e.g., via panel models), as our findings suggest that EF may partially mediate the relation between math in preschool and math in kindergarten.

Fourth, it is important to note that there was attrition across the four waves of data, particularly as children were transitioning from preschool to kindergarten (between Times 2 and 3). Although we accounted for missing data by using robust maximum likelihood and included Head Start status in all of our models (which predicted missingness between these waves), different patterns of reciprocal relations in preschool and kindergarten may be due to attrition.

Finally, although our sample was diverse in terms of socioeconomic status, it was less ethnically diverse. We relied on a convenience sample for the present analyses, and future research is needed to replicate our findings with more representative and



ethnically diverse samples to determine whether or not the findings generalize to other populations.

## Conclusions

Findings from this study have potential implications for instruction and intervention development that need to be investigated in a more targeted manner. It may be important to consider the EF demands on mathematical instruction at these ages. The relation between EF and math may be something that can be capitalized on through instruction. Integrating the domains at a very targeted level (e.g., that includes appropriate individual scaffolding) may be a useful mechanism for enhancing success across domains. Further, intervention efforts focused on EF (or math) may also have a beneficial effect on children's math (or EF) development. Although our analyses preclude causality, the bidirectional associations, as well as correlated growth trajectories, between EF and math suggests that interventions and programs that contain both EF and academic training, particularly in math, may be a potential avenue for affecting change during the transition to kindergarten. Future research examining causal connections between these domains at a more nuanced level is needed.

Findings from this study also suggest that, without intervention, children's relative standing on EF, math, and literacy assessments are fairly stable over time. This finding has implications for future theoretical work examining the development of these constructs. More research is needed to identify predictors of these skills prior to and during preschool at the biological, familial, and socioeconomic levels.

In sum, the current study replicates and extends current literature exploring EF, math, and literacy. Unlike previous work, we used a multi-analytic approach and found converging evidence for the longitudinal relations between EF and math and weaker relations between EF and literacy. These findings expand upon what was found in the study conducted by Fuhs and colleagues (2014). With the addition of a fourth time point at the beginning of kindergarten, we were able to contribute to current research by improving the specificity of the relations between EF and academic skills by identifying at which points the relations change during the transition to kindergarten at a more fine-grained level. Changes in these relations may be due to factors within the preschool and kindergarten classrooms, such as instructional methods and alignment to children's needs, or due to the constructs being assessed at those ages. This change in relation is important for the development of instructional strategies and interventions that aim to improve either EF or math. In preschool, it may be more efficacious to target both EF and math simultaneously, whereas in kindergarten, targeting EF as a foundational skill set may be more important. Alternatively, there may be differential relations between aspects of EF and mathematics where EF is only related to certain mathematics skills (Lan et al., 2011; Purpura, Schmitt, & Ganley, 2017; Purpura & Ganley, 2014) that affect this relation. These differential relations may need to be accounted for in intervention and curricular development. Nonetheless, findings from both sets of analyses suggest that fostering the development of EF and early math skills during the transition to kindergarten may be a potentially important avenue for promoting school readiness and fostering academic success that needs to be investigated more thoroughly.

## References

- Barbareis, W. J., Katusic, S. K., Colligan, R. C., Weaver, A. L., & Jacobsen, S. J. (2005). Math learning disorder: Incidence in a population-based birth cohort, 1976–82, Rochester, Minn. *Ambulatory Pediatrics*, 5, 281–289. <http://dx.doi.org/10.1367/A04-209R.1>
- Blackwell, K. A., Cepeda, N. J., & Munakata, Y. (2009). When simple things are meaningful: Working memory strength predicts children's cognitive flexibility. *Journal of Experimental Child Psychology*, 103, 241–249. <http://dx.doi.org/10.1016/j.jecp.2009.01.002>
- Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences*, 29, 109–125.
- Blair, C., & Raver, C. C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology*, 66, 711–731. <http://dx.doi.org/10.1146/annurev-psych-010814-015221>
- Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, 78, 647–663. <http://dx.doi.org/10.1111/j.1467-8624.2007.01019.x>
- Blair, C., Ursache, A., Greenberg, M., & Vernon-Feagans, L. (2015). Multiple aspects of self-regulation uniquely predict mathematics but not letter-word knowledge in the early elementary grades. *Developmental Psychology*, 51, 459–472. <http://dx.doi.org/10.1037/a0038813>
- Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology*, 33, 205–228. <http://dx.doi.org/10.1080/87565640801982312>
- Bull, R., Johnston, R. S., & Roy, J. A. (1999). Exploring the roles of the visual-spatial sketch pad and central executive in children's arithmetical skills: Views from cognition and developmental neuropsychology. *Developmental Neuropsychology*, 15, 421–442. <http://dx.doi.org/10.1080/87565649909540759>
- Burrage, M. S., Cameron Ponitz, C., McCready, E. A., Shah, P., Sims, B. C., Jewkes, A. M., & Morrison, F. J. (2008). Age- and schooling-related effects on executive functions in young children: A natural experiment. *Child Neuropsychology*, 14, 510–524. <http://dx.doi.org/10.1080/09297040701756917>
- Cameron Ponitz, C., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. *Developmental Psychology*, 45, 605–619. <http://dx.doi.org/10.1037/a0015365>
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105. <http://dx.doi.org/10.1037/h0046016>
- Cannon, J., & Ginsburg, H. P. (2008). 'Doing the math': Maternal beliefs about early mathematics versus language learning. *Early Education and Development*, 19, 238–260. <http://dx.doi.org/10.1080/10409280801963913>
- Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology*, 28, 595–616. [http://dx.doi.org/10.1207/s15326942dn2802\\_3](http://dx.doi.org/10.1207/s15326942dn2802_3)
- Chen, Q., Hughes, J. N., & Kwok, O. M. (2014). Differential growth trajectories for achievement among children retained in first grade: A growth mixture model. *The Elementary School Journal*, 114, 327–353. <http://dx.doi.org/10.1086/674054>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, 9, 233–255. [http://dx.doi.org/10.1207/S15328007SEM0902\\_5](http://dx.doi.org/10.1207/S15328007SEM0902_5)
- Christian, K., Bachman, H. J., & Morrison, F. J. (2001). Schooling and



- cognitive development. In R. J. Sternberg & E. L. Grigorenko (Eds.), *Environmental effects on cognitive abilities* (pp. 287–335). Mahwah, NJ: Erlbaum.
- Clark, C. A. C., Pritchard, V. E., & Woodward, L. J. (2010). Preschool executive functioning abilities predict early mathematics achievement. *Developmental Psychology, 46*, 1176–1191. <http://dx.doi.org/10.1037/a0019672>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. <http://dx.doi.org/10.1007/BF02310555>
- Dowsett, S. M., & Livesey, D. J. (2000). The development of inhibitory control in preschool children: Effects of “executive skills” training. *Developmental Psychobiology, 36*, 161–174. [http://dx.doi.org/10.1002/\(SICI\)1098-2302\(200003\)36:2<161::AID-DEV7>3.0.CO;2-0](http://dx.doi.org/10.1002/(SICI)1098-2302(200003)36:2<161::AID-DEV7>3.0.CO;2-0)
- DeFlorio, L., & Beliakoff, A. (2015). Socioeconomic status and preschoolers’ mathematical knowledge: The contribution of home activities and parent beliefs. *Early Education and Development, 26*, 319–341. <http://dx.doi.org/10.1080/10409289.2015.968239>
- Duckworth, A. L., Tsukayama, E., & May, H. (2010). Establishing causality using longitudinal hierarchical linear modeling: An illustration predicting achievement from self-control. *Social Psychological and Personality Science, 1*, 311–317. <http://dx.doi.org/10.1177/1948550609359707>
- Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428–1446. <http://dx.doi.org/10.1037/0012-1649.43.6.1428>
- Espy, K. A., McDiarmid, M. M., Cwik, M. F., Stalets, M. M., Hamby, A., & Senn, T. E. (2004). The contribution of executive functions to emergent mathematic skills in preschool children. *Developmental Neuropsychology, 26*, 465–486. [http://dx.doi.org/10.1207/s15326942dn2601\\_6](http://dx.doi.org/10.1207/s15326942dn2601_6)
- Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development, 10*, 483–527. [http://dx.doi.org/10.1016/0885-2014\(95\)90024-1](http://dx.doi.org/10.1016/0885-2014(95)90024-1)
- Fuhs, M. W., Nesbitt, K. T., Farran, D. C., & Dong, N. (2014). Longitudinal associations between executive functioning and academic skills across content areas. *Developmental Psychology, 50*, 1698–1709. <http://dx.doi.org/10.1037/a0036633>
- Gathercole, S. E., Pickering, S. J., Knight, C., & Stegmann, Z. (2004). Working memory skills and educational attainment: Evidence from national curriculum assessments at 7 and 14 years of age. *Applied Cognitive Psychology, 18*, 1–16. <http://dx.doi.org/10.1002/acp.934>
- Geldhof, G. J., Pornprasertmanit, S., Schoemann, A. M., & Little, T. D. (2013). Orthogonalizing through residual centering: Extended applications and caveats. *Educational and Psychological Measurement, 73*, 27–46. <http://dx.doi.org/10.1177/0013164412445473>
- Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*, 72–91. <http://dx.doi.org/10.1037/a0032138>
- Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report—Giving Child and Youth Development Knowledge Away, 22*, 1–24.
- Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*, 255–274. <http://dx.doi.org/10.3102/01623737011003255>
- Hancock, G. R., Kuo, W., & Lawrence, F. R. (2001). An illustration of second-order growth models. *Structural Equation Modeling, 8*, 470–489. [http://dx.doi.org/10.1207/S15328007SEM0803\\_7](http://dx.doi.org/10.1207/S15328007SEM0803_7)
- Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20*, 102–116. <http://dx.doi.org/10.1037/a0038889>
- Hassinger-Das, B., Jordan, N. C., Glutting, J., Irwin, C., & Dyson, N. (2014). Domain-general mediators of the relation between kindergarten number sense and first-grade mathematics achievement. *Journal of Experimental Child Psychology, 118*, 78–92. <http://dx.doi.org/10.1016/j.jecp.2013.09.008>
- Hongwanishkul, D., Happaney, K. R., Lee, W. S. C., & Zelazo, P. D. (2005). Assessment of hot and cool executive function in young children: Age-related changes and individual differences. *Developmental Neuropsychology, 28*, 617–644. [http://dx.doi.org/10.1207/s15326942dn2802\\_4](http://dx.doi.org/10.1207/s15326942dn2802_4)
- Hughes, C., Ensor, R., Wilson, A., & Graham, A. (2009). Tracking executive function across the transition to school: A latent variable approach. *Developmental Neuropsychology, 35*, 20–36. <http://dx.doi.org/10.1080/87565640903325691>
- Huizinga, M., Dolan, C. V., & van der Molen, M. W. (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia, 44*, 2017–2036. <http://dx.doi.org/10.1016/j.neuropsychologia.2006.01.010>
- Klingberg, T. (2006). Development of a superior frontal – intraparietal network for visuo-spatial working memory. *Neuropsychologia, 44*, 2171–2177.
- Lan, X., Legare, C. H., Cameron Ponitz, C., Li, S., & Morrison, F. J. (2011). Investigating the links between the subcomponents of executive function and academic achievement: A cross-cultural analysis of Chinese and American preschoolers. *Journal of Experimental Child Psychology, 108*, 677–692. <http://dx.doi.org/10.1016/j.jecp.2010.11.001>
- La Paro, K. M., & Pianta, R. C. (2000). Predicting children’s competence in the early school years. A meta-analytic review. *Review of Educational Research, 70*, 443–484. <http://dx.doi.org/10.3102/00346543070004443>
- Layzer, J. L., Goodsen, B. D., & Moss, M. (1993). *Life in preschool: Vol. 1. Observational study of early childhood programs for disadvantaged four-year-olds*. Cambridge, MA: Abt Associates.
- LeFevre, J. A., Fast, L., Skwarchuk, S. L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development, 81*, 1753–1767. <http://dx.doi.org/10.1111/j.1467-8624.2010.01508.x>
- Lehto, J. E., Juujärvi, P., Kooistra, L., & Pulkkinen, L. (2003). Dimensions of executive function: Evidence from children. *British Journal of Developmental Psychology, 21*, 59–80. <http://dx.doi.org/10.1348/026151003321164627>
- Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53–76. [http://dx.doi.org/10.1207/s15327906mbr3201\\_3](http://dx.doi.org/10.1207/s15327906mbr3201_3)
- McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 561–614). New York, NY: Springer. [http://dx.doi.org/10.1007/978-1-4613-0893-5\\_17](http://dx.doi.org/10.1007/978-1-4613-0893-5_17)
- McClelland, M. M., Acock, A. C., & Morrison, F. J. (2006). The impact of kindergarten learning-related skills on academic trajectories at the end of elementary school. *Early Childhood Research Quarterly, 21*, 471–490. <http://dx.doi.org/10.1016/j.ecresq.2006.09.003>
- McClelland, M. M., Acock, A. C., Piccinin, A., Rhea, S. A., & Stallings, M. C. (2013). Relations between preschool attention span-persistence and age 25 educational outcomes. *Early Childhood Research Quarterly, 28*, 314–324. <http://dx.doi.org/10.1016/j.ecresq.2012.07.008>
- McClelland, M. M., & Cameron, C. E. (2012). Self-Regulation in early childhood: Improving conceptual clarity and developing ecologically valid measures. *Child Development Perspectives, 6*, 136–142. <http://dx.doi.org/10.1111/j.1750-8606.2011.00191.x>
- McClelland, M. M., Cameron, C. E., Connor, C. M., Farris, C. L., Jewkes, A. M., & Morrison, F. J. (2007). Links between behavioral regulation and preschoolers’ literacy, vocabulary, and math skills. *Developmental Psychology, 43*, 947–959. <http://dx.doi.org/10.1037/0012-1649.43.4.947>
- McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., & Pratt, M. E. (2014). Predictors of early growth in academic achievement: The head-toes-knees-shoulders task. *Frontiers*



- in *Psychology*, 5, 599. Advance online publication. <http://dx.doi.org/10.3389/fpsyg.2014.00599>
- McClelland, M. M., Cameron, C. E., Wanless, S. B., & Murray, A. (2007). Executive function, behavioral self-regulation, and social-emotional competence: Links to school readiness. In O. N. Saracho & B. Spodek (Eds.), *Contemporary perspectives on social learning in early childhood education* (pp. 83–107). Charlotte, NC: Information Age.
- McClelland, M. M., Geldof, J., Cameron, C. E., & Wanless, S. B. (2015). Development and self-regulation. In W. F. Overton & P. C. M. Molenaar (Eds.), *Theory and Method. Vol. 1 of the Handbook of child psychology and developmental science* (7th ed.). Hoboken, NJ: Wiley. Advance online publication. <http://dx.doi.org/10.1002/9781118963418.childpsy114>
- McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, 23, 1–21. <http://dx.doi.org/10.1111/j.2044-8317.1970.tb00432.x>
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41, 49–100. <http://dx.doi.org/10.1006/cogp.1999.0734>
- Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., . . . Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 2693–2698. <http://dx.doi.org/10.1073/pnas.1010076108>
- Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S., & Mather, N. (2005a). *Bateria III Woodcock–Muñoz: Pruebas de aprovechamiento* [Tests of Achievement]. Itasca, IL: Riverside.
- Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S., & Mather, N. (2005b). *Bateria III Woodcock–Muñoz: Pruebas de habilidades cognitivas* [Tests of Cognitive Abilities]. Itasca, IL: Riverside.
- Muthén, B. O. (2011, May 19). *Factor scores as dependent variables* [Msg. 2]. Message posted to <http://www.statmodel.com/discussion/messages/9/7397.html?1359828921>
- Muthén, L. K. (2010, April 25). *Convergence* [Msg. 21]. Message posted to <http://www.statmodel.com/discussion/messages/12/17.html?1438886834>
- Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.
- National Early Literacy Panel. (2008). *Report of the National Early Literacy Panel: Developing early literacy*. Washington, DC: U. S. Department of Health and Human Services.
- National Mathematics Advisory Panel. (2008). *Foundations for success: The Final report of the National Mathematics Advisory Panel*. Washington, DC: U. S. Department of Education.
- Purpura, D. J., & Ganley, C. M. (2014). Working memory and language: Skill-specific or domain-general relations to mathematics? *Journal of Experimental Child Psychology*, 122, 104–121. <http://dx.doi.org/10.1016/j.jecp.2013.12.009>
- Purpura, D. J., Hume, L. E., Sims, D. M., & Lonigan, C. J. (2011). Early literacy and early numeracy: The value of including early literacy skills in the prediction of numeracy development. *Journal of Experimental Child Psychology*, 110, 647–658. <http://dx.doi.org/10.1016/j.jecp.2011.07.004>
- Purpura, D. J., Schmitt, S. A., & Ganley, C. M. (2017). Foundations of mathematics and literacy: The role of executive functioning components. *Journal of Experimental Child Psychology*, 153, 15–34. <http://dx.doi.org/10.1016/j.jecp.2016.08.010>
- Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21, 173–184. <http://dx.doi.org/10.1177/01466216970212006>
- Richland, L. E., & Burchinal, M. R. (2013). Early executive function predicts reasoning development. *Psychological Science*, 24, 87–92. <http://dx.doi.org/10.1177/0956797612450883>
- Rueda, M. R., Posner, M. I., & Rothbart, M. K. (2005). The development of executive attention: Contributions to the emergence of self-regulation. *Developmental Neuropsychology*, 28, 573–594. [http://dx.doi.org/10.1207/s15326942dn2802\\_2](http://dx.doi.org/10.1207/s15326942dn2802_2)
- Schmitt, S. A., Pratt, M., & McClelland, M. M. (2014). Examining the validity of behavioral self-regulation tools in predicting preschoolers' academic achievement. *Early Education and Development*, 25, 641–660. <http://dx.doi.org/10.1080/10409289.2014.850397>
- Skibbe, L. E., Hindman, A. H., Connor, C. M., Housey, M., & Morrison, F. J. (2013). Relative contributions of pre-kindergarten and kindergarten to children's literacy and mathematics skills. *Early Education and Development*, 24, 687–703. <http://dx.doi.org/10.1080/10409289.2012.712888>
- Skibbe, L. E., Phillips, B. M., Day, S. L., Brophy-Herb, H. E., & Connor, C. M. (2012). Children's early literacy growth in relation to classmates' self-regulation. *Journal of Educational Psychology*, 104, 541–553. <http://dx.doi.org/10.1037/a0029153>
- Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, 66, 563–575. <http://dx.doi.org/10.1007/BF02296196>
- Skwarchuk, S. L., Sowinski, C., & LeFevre, J. A. (2014). Formal and informal home learning activities in relation to children's early numeracy and literacy skills: The development of a home numeracy model. *Journal of Experimental Child Psychology*, 121, 63–84. <http://dx.doi.org/10.1016/j.jecp.2013.11.006>
- Stevenson, H. W., & Newman, R. S. (1986). Long-term prediction of achievement and attitudes in mathematics and reading. *Child Development*, 57, 646–659. <http://dx.doi.org/10.2307/1130343>
- Strommen, E. A. (1973). Verbal self-regulation in a children's game: Impulsive errors on “Simon Says.” *Child Development*, 849–853. Advance online publication.
- Symonds, J. E., & Gorard, S. (2010). Death of mixed methods? Or the rebirth of research as a craft. *Evaluation and Research in Education*, 23, 121–136. <http://dx.doi.org/10.1080/09500790.2010.483514>
- Wanless, S. B., McClelland, M. M., Acock, A. C., Ponitz, C. C., Son, S.-H., Lan, X., . . . Li, S. (2011). Measuring behavioral regulation in four societies. *Psychological Assessment*, 23, 364–378. <http://dx.doi.org/10.1037/a0021768>
- Welsh, J. A., Nix, R. L., Blair, C., Bierman, K. L., & Nelson, K. E. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of Educational Psychology*, 102, 43–53. <http://dx.doi.org/10.1037/a0016738>
- Werts, C. E., Linn, R. L., & Jöreskog, K. G. (1974). Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement*, 34, 25–33. <http://dx.doi.org/10.1177/001316447403400104>
- Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development*, 69, 848–872. <http://dx.doi.org/10.1111/j.1467-8624.1998.tb06247.x>
- Wiebe, S. A., Espy, K. A., & Charak, D. (2008). Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. *Developmental Psychology*, 44, 575–587. <http://dx.doi.org/10.1037/0012-1649.44.2.575>
- Willoughby, M. T., Blair, C. B., Wirth, R. J., & Greenberg, M. (2012). The measurement of executive function at age 5: Psychometric properties and relationship to academic achievement. *Psychological Assessment*, 24, 226–239. <http://dx.doi.org/10.1037/a0025361>
- Willoughby, M. T., Kupersmidt, J. B., & Voegler-Lee, M. E. (2012). Is preschool executive function causally related to academic achievement? *Child Neuropsychology*, 18, 79–91. <http://dx.doi.org/10.1080/09297049.2011.578572>
- Willoughby, M. T., Pek, J., & Blair, C. B. (2013). Measuring executive function in early childhood: A focus on maximal reliability and the

- derivation of short forms. *Psychological Assessment*, 25, 664–670. <http://dx.doi.org/10.1037/a0031747>
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001a). *Woodcock–Johnson III Tests of Achievement*. Itasca, IL: Riverside.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001b). *Woodcock–Johnson III Tests of Cognitive Abilities*. Itasca, IL: Riverside.
- Woodcock, R. W., & Muñoz-Sandoval, A. F. (1993). An IRT approach to cross-language test equating and interpretation. *European Journal of Psychological Measurement*, 32, 233–241.
- Wu, W., Selig, J. P., & Little, T. D. (2013). Longitudinal models. In T. D. Little (Ed.), *Oxford handbook of quantitative methods* (pp. 387–410). New York, NY: Oxford University Press.
- Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, 1, 297–301. <http://dx.doi.org/10.1038/nprot.2006.46>
- Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). II. National Institutes of Health Toolbox Cognition Battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development*, 78, 16–33. <http://dx.doi.org/10.1111/mono.12032>
- Zelazo, P. D., & Ulrich, M. (2011). Executive function in typical and atypical development. In U. Goswami (Ed.), *The Wiley–Blackwell handbook of childhood cognitive development* (2nd ed., pp. 574–603). Hoboken, NJ: Wiley-Blackwell.



## Appendix A

## SAS Code for Dividing Items by Constants

```

DATA use; SET use;

/*DIVIDE HTKS SUMS BY 15*/
htks1 = sum (of htkss1_1 htkss2_1 htkss3_1)/15;
htks2 = sum (of htkss1_2 htkss2_2 htkss3_2)/15;
htks3 = sum (of htkss1_3 htkss2_3 htkss3_3)/15;
htks4 = sum (of htkss1_4 htkss2_4 htkss3_4)/15;

/*DIVIDE DCCS SUMS BY 3*/
dccs1 = sum (of dccs1_1 dccs2_1 dccs3_1 dccs4_1)/3;
dccs2 = sum (of dccs1_2 dccs2_2 dccs3_2 dccs4_2)/3;
dccs3 = sum (of dccs1_3 dccs2_3 dccs3_3 dccs4_3)/3;
dccs4 = sum (of dccs1_4 dccs2_4 dccs3_4 dccs4_4)/3;

/*DIVIDE ALL WJ SCORES BY 10*/
wjapw_1 = wjapw_1/10;
wjapw_2 = wjapw_2/10;
wjapw_3 = wjapw_3/10;
wjapw_4 = wjapw_4/10;
wjllw_1 = wjllw_1/10;
wjllw_2 = wjllw_2/10;
wjllw_3 = wjllw_3/10;
wjllw_4 = wjllw_4/10;
wjpvw_1 = wjpvw_1/10;
wjpvw_2 = wjpvw_2/10;
wjpvw_3 = wjpvw_3/10;
wjpvw_4 = wjpvw_4/10;
wjwmw_1 = wjwmw_1/10;
wjwmw_2 = wjwmw_2/10;
wjwmw_3 = wjwmw_3/10;
wjwmw_4 = wjwmw_4/10;

/*RECENTER AGE AT 4.5 YRS*/
ageyrs_1 = ageyrs_1-4.5;
ageyrs_2 = ageyrs_2-4.5;
ageyrs_3 = ageyrs_3-4.5;
ageyrs_4 = ageyrs_4-4.5;

/*DROP INDIVIDUAL ITEMS, ANALYSIS AT COMPOSITE LEVEL ONLY*/
drop htkss1_1 htkss1_2 htkss1_3 htkss1_4
      htkss2_1 htkss2_2 htkss2_3 htkss2_4
      htkss3_1 htkss3_2 htkss3_3 htkss3_4
      dccs1_1 dccs2_1 dccs3_1 dccs4_1
      dccs1_2 dccs2_2 dccs3_2 dccs4_2
      dccs1_3 dccs2_3 dccs3_3 dccs4_3
      dccs1_4 dccs2_4 dccs3_4 dccs4_4
      hstart_2 cspan_2 cspan_3 cspan_1;
RUN;

```

(Appendices continue)

Appendix B  
Additional Tables and Figures

Table B1  
*Residual Correlations from Final Structural Equation Model*

Construct	1	2	3	4	5	6	7	8	9	10	11	12
1. EF1	1.00											
2. Math1	0.66***	4.01										
3. Literacy1	0.45***	0.40***	5.46									
4. EF2	—	—	—	.28								
5. Math2	—	—	—	00.25*	1.43							
6. Literacy2	—	—	—	0.11	0.10	2.20						
7. EF3	—	—	—	—	—	—	0.29					
8. Math3	—	—	—	—	—	—	0	1.11				
9. Literacy3	—	—	—	—	—	—	0.05	0.16**	2.43			
10. EF4	—	—	—	—	—	—	—	—	—	0.40		
11. Math4	—	—	—	—	—	—	—	—	—	0.13	1.01	
12. Literacy4	—	—	—	—	—	—	—	—	—	0.06	0.06	3.76

*Note.* A dash indicates values were not estimated. Variances and residual variances on diagonal, correlations below diagonal. Squared factor loadings (therefore representing item variances) provided for math and literacy. Indicators were divided by constants to make their variances more homogenous, thus expediting model convergence (e.g., Muthén, 2010). Dashes indicate that the data were not obtained/reported. EF = executive functioning.  
\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

Table B2  
*Comparison of BIC for Final Panel Model and Growth Curve (N = 424)*

Model	BIC
Panel model	30,624.303
Growth curve	30,435.914

*Note.* BIC = Bayesian information criterion.

(Appendices continue)



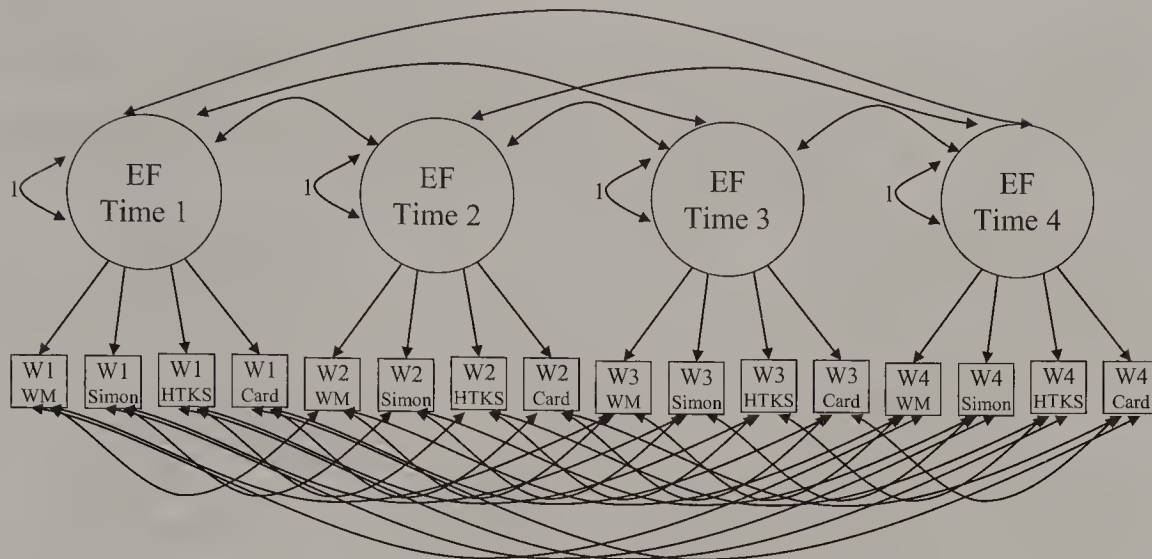


Figure B1. Path diagram representing the executive functioning (EF) component of the initial confirmatory factor analysis (CFA). Mean structure and indicator residuals are omitted from the diagram, but all indicator residuals and intercepts were freely estimated. All latent means were fixed to zero. All indicators were controlled for covariates (not shown).

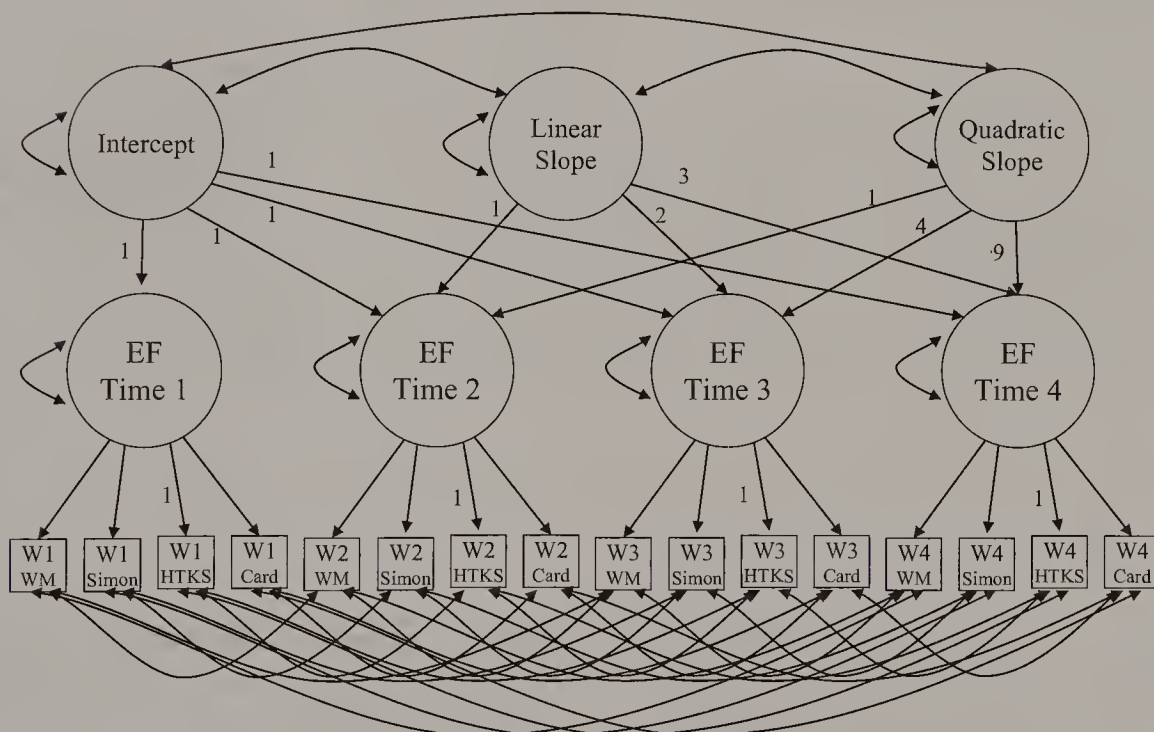


Figure B2. Path diagram representing the executive functioning (EF) component of the initial latent growth curve modeling (LGCM). Mean structure and indicator residuals are omitted from the diagram, but all indicator residuals were freely estimated, with factor loadings and indicator intercepts estimated but equated across time. Head-Toes-Knees-Shoulders task (HTKS) served as a marker variable, with its loading fixed to 1.00 and intercept fixed to 0.00. All latent means for EF were fixed to zero and means for all growth parameters (intercept and two slopes) were freely estimated. All growth parameters were controlled for covariates (not shown).

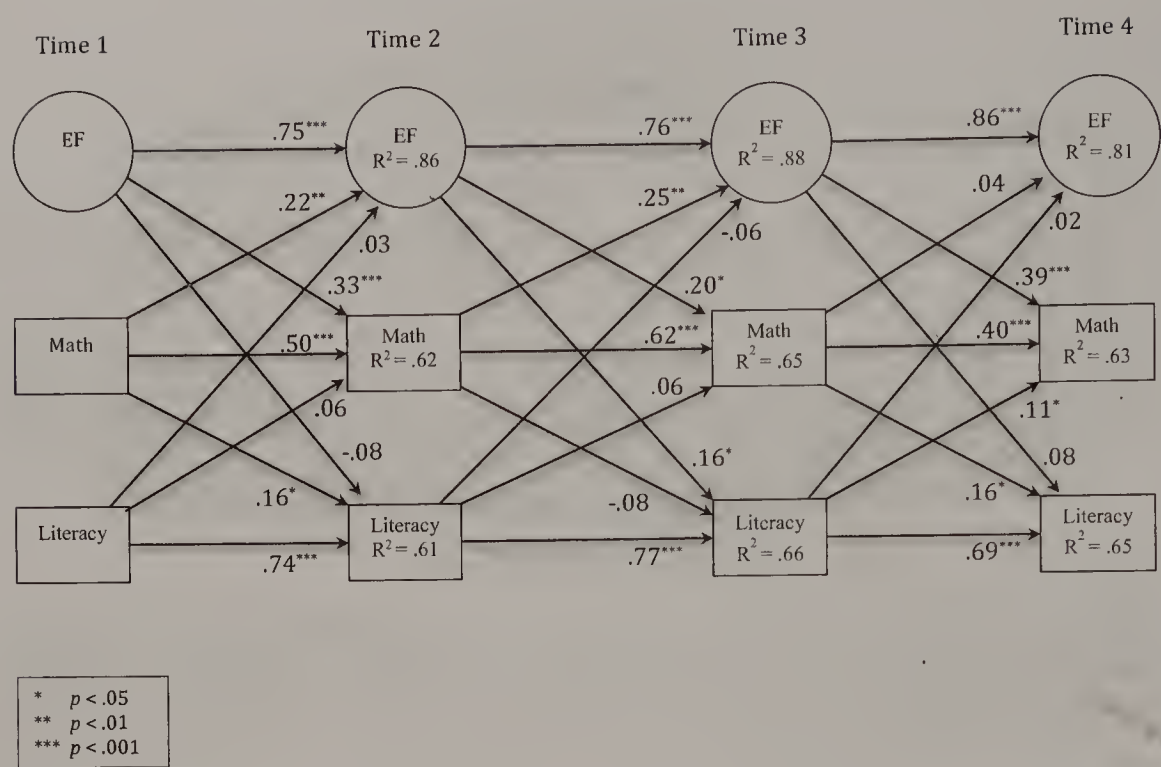


Figure B3. Path diagram for the final structural model (standardized coefficients). Variances, residual variances, and within-wave covariances provided in Table B1. Time 1 = fall of preschool; Time 2 = spring of preschool; Time 3 = fall of kindergarten; Time 4 = spring of kindergarten.

Received July 23, 2015  
Revision received January 18, 2017  
Accepted January 19, 2017 ■



# Achievement Goals, Reasons for Goal Pursuit, and Achievement Goal Complexes as Predictors of Beneficial Outcomes: Is the Influence of Goals Reducible to Reasons?

Nicolas Sommet

University of Rochester and University of Lausanne

Andrew J. Elliot

University of Rochester

In the present research, we proposed a systematic approach to disentangling the shared and unique variance explained by achievement goals, reasons for goal pursuit, and specific goal-reason combinations (i.e., achievement goal complexes). Four studies using this approach (involving nearly 1,800 participants) led to 3 basic sets of findings. First, when testing goals and reasons *separately*, mastery (-approach) goals and autonomous reasons explained variance in beneficial experiential (interest, satisfaction, positive emotion) and self-regulated learning (deep learning, help-seeking, challenging tasks, persistence) outcomes. Second, when testing goals and reasons *simultaneously*, mastery goals and autonomous reasons explained independent variance in most of the outcomes, with the predictive strength of each being diminished. Third, when testing goals, reasons, and goal complexes *together*, the autonomous mastery goal complex explained incremental variance in most of the outcomes, with the predictive strength of both mastery goals and autonomous reasons being diminished. Comparable results were observed for performance (-approach) goals, the autonomous performance goal complex, and performance goal-relevant outcomes. These findings suggest that achievement goals and reasons are both distinct and overlapping constructs, and that neither unilaterally eliminates the influence of the other. Integrating achievement goals and reasons offers the most promising avenue for a full account of competence motivation.

## *Educational Impact and Implications Statement*

The present research seeks to disentangle the influence of “what” individuals want to achieve (type of goals), “why” they want to achieve (type of reasons), and specific “what” and “why” combinations (type of goal-reason combinations). In four studies, we showed that mastery goals (striving for task mastery), autonomous reasons (striving because it is stimulating and valued), and a specific mastery goal—autonomous reason combination (striving for task mastery because it is stimulating and valued) all made separate positive contributions to beneficial achievement-relevant outcomes (e.g., interest, positive emotion, deep learning). Comparable results were observed for performance goals (striving to outperform others) and a specific performance goal—autonomous reason combination (striving to outperform others because it is stimulating and valuable). The present findings indicate that both type of goals and type of reasons are important for a full understanding of achievement motivation.

**Keywords:** achievement goal, autonomous and controlled reasons, self-determination theory, achievement goal complex

The achievement goal approach provides a framework for understanding the direction of behavior, addressing the question of *what* individuals want to achieve (Dweck, 1986; Maehr & Nicholls, 1980; Nicholls, 1984). However, a complete conceptual framework of

achievement motivation must also account for the energization of behavior, addressing the question of *why* individuals want to achieve (Elliot & Thrash, 2001).

The “whys” (i.e., reasons) behind achievement goals can be conceptualized in many ways (e.g., social values, achievement motives, Dompnier, Darnon, & Butera, 2009; McClelland, 1985). However, in recent years researchers have focused mostly on reasons derived from self-determination theory (SDT, Ryan & Deci, 2000). In several studies, researchers have reported that the influence of achievement goals on beneficial outcomes is no longer statistically significant when partialing out the variance explained by the SDT-derived reasons connected with the achievement goals (for a review, see Vansteenkiste, Lens, Elliot, Soenens, & Mouratidis, 2014). These findings are sometimes interpreted as indicating that the influence of achievement goals is reducible to the reasons behind them, thereby questioning the importance of achievement goals in the study of motivation.

This article was published Online First March 30, 2017.

Nicolas Sommet, Department of Clinical and Social Sciences in Psychology, University of Rochester and LINES, SSP, ISS, University of Lausanne; Andrew J. Elliot, Department of Clinical and Social Sciences in Psychology, University of Rochester.

This research was supported by a postdoctoral UNIL/CHUV fellowship (University of Lausanne, Switzerland) awarded to the first author.

Correspondence concerning this article should be addressed to Nicolas Sommet, LINES, SSP, ISS, University of Lausanne Bâtiment Géopolis, Bureau #5785, Quartier UNIL-Mouline, Switzerland. E-mail: nicolas.sommet@unil.ch

In the present research, we take a step back to carefully examine this empirical work and to reconsider the conclusions that can be drawn from it. We propose a systematic approach for studying achievement goals, reasons, and specific achievement goal-reason combinations (i.e., achievement goal complexes; Elliot & Thrash, 2001). We use this approach in four studies to disentangle the shared and unique variance explained by these motivational constructs in predicting the most commonly investigated beneficial outcomes in the achievement domain. We believe that this approach holds considerable promise, in that it demonstrates how achievement goals fit in a broader theory of achievement motivation.

### Mastery Goals as a Predictor of Beneficial Outcomes

Achievement goals are social-cognitive mental foci that direct individuals' responses in competence-relevant situations (Elliot, 1999). Achievement goal researchers focus primarily on two types of competence-based goals, crossed by the approach-avoidance distinction (for a historical review, see Elliot, 2005). Mastery-focused individuals use a task- or self-referenced standard in competence evaluation, whereas performance-focused individuals use an other-referenced standard. Both mastery and performance goals involve striving to approach competence or avoid incompetence, resulting in a  $2 \times 2$  model of achievement goals: mastery-approach, mastery-avoidance, performance-approach, and performance-avoidance.

In the literature, mastery-approach goals are primarily linked to a pattern of adaptive outcomes, performance-approach goals to a mixed pattern of adaptive and maladaptive outcomes, and the two avoidance goals to varied patterns of maladaptive outcomes (for meta-analyses, see Baranik, Stanley, Bynum, & Lance, 2010; Huang, 2011, 2016; Hulleman, Schrager, Bodmann, & Harackiewicz, 2010; Van Yperen, Blaga, & Postmes, 2014, 2015). In the present research, we are interested in separating the influence of achievement goals from the influence of reasons when predicting beneficial achievement-relevant outcomes. It is therefore critical to select goals and reasons that are clearly adaptive (and whose beneficial influences are comparable in nature and scope). Accordingly, our primary focus is on mastery-approach goals (i.e., mastering a task, improving over time; hereafter referred to as mastery goals), although in our final study we extend the focus to performance-approach goals (i.e., outperforming others; hereafter referred to as performance goals).

Two types of adaptive achievement-relevant outcomes are reliably associated with mastery goals. First, mastery goals are positively related to beneficial *experiential* outcomes, that is, positive affective and phenomenological responses to achievement tasks (Harackiewicz, Barron, Carter, Lehto, & Elliot, 1997; Pekrun, 2006). Mastery goals are thought to direct attention to the achievement activity itself and increase appraisals of task controllability and self-efficacy, thereby facilitating the positive subjective value of the task (Dweck, 1999; Kaplan & Maehr, 2007; Pekrun, Elliot, & Maier, 2006). For instance, in the workplace, mastery goals have been shown to positively predict job interest (Retelsdorf, Butler, Streblow, & Schiefele, 2010), job satisfaction (Janssen & Van Yperen, 2004), and job positive emotion (Fisher, Minbashian, Beckmann, & Wood, 2013). Second, mastery goals are positively related to beneficial *self-regulated learning* outcomes, that is, metacognitive, strategic, proactive responses to achievement tasks (Pintrich, 1999; Zimmerman, 1989). Mastery goals require the attainment of task-focused and intrapersonal

standards, which promote a fully engaged approach to learning and full effort expenditure (Meece, Anderman, & Anderman, 2006; Nicholls, 1989; Senko, Hama, & Belmonte, 2013). As such, mastery goals have been shown to positively predict deep-processing (Diseth, 2011), interpersonal help-seeking behavior (Karabenick, 2004), a preference for challenging tasks (Ames & Archer, 1988), and task persistence (Sideridis & Kaplan, 2011).

### Autonomous Reasons as a Predictor of Beneficial Outcomes

SDT is a theory of motivation that highlights the importance of underlying reasons for behavior, including goal-directed behavior (Deci & Ryan, 2000; Sheldon, 2004). The theory distinguishes between two primary types of reasons for goal pursuit. Autonomous reasons include pursuing goals because they are fun or enjoyable (intrinsic regulation), or because one identifies with them as important or meaningful (identified regulation); controlled reasons include pursuing goals because they enable one to bolster the ego or avoid feeling shame (introjected regulation), or because they allow one to obtain a reward (external regulation; Deci & Ryan, 2000). In the literature, autonomous reasons are most commonly predictors of beneficial outcomes, whereas controlled reasons are most commonly predictors of detrimental outcomes (Ratelle, Guay, Vallerand, Larose, & Senécal, 2007). Accordingly, our primary focus is on autonomous reasons (although in all of our studies we assessed and controlled for controlled reasons, as well).

Autonomous reasons for goal pursuit are associated with the same beneficial outcomes as those reviewed above for mastery goals (for a review, see Ryan & Deci, 2006). First, autonomous reasons are positively related to beneficial *experiential* outcomes, because they involve acting in a more volitional way, thereby making the activity more enjoyable and immersive (Vansteenkiste, Lens, et al., 2014). For instance, in the workplace, autonomous reasons have been shown to positively predict job interest (Gagné & Deci, 2005), job satisfaction (Lam & Gurland, 2008), and job positive emotion (Gagné et al., 2010). Second, autonomous reasons are positively related to beneficial *self-regulated learning* outcomes, because goal pursuit is viewed as a positive challenge, providing a meaningful impetus for effort expenditure and personal growth (Deci, Vallerand, Pelletier, & Ryan, 1991). Specifically, empirical work has shown that these reasons positively predict deep learning strategy (Vansteenkiste, Zhou, Lens, & Soenens, 2005), interpersonal help-seeking behavior (Skaalvik & Skaalvik, 2013), a preference for challenge (Standage, Duda, & Ntoumanis, 2005), and persistence (Vallerand, Fortier, & Guay, 1997).

### Combining Mastery Goals and Autonomous Reasons as Predictors of Beneficial Outcomes

Any given achievement goal may be adopted for a variety of reasons. These reasons may vary from competence-relevant (e.g., to succeed at university; Dompnier et al., 2009) to not competence-relevant (e.g., to gain respect from others; Urdañ & Mestas, 2006), and from intrapersonally evoked (e.g., a desire to experience pride; Urdañ, 2004a) to environmentally evoked (e.g., a teacher demand; Wolters, 2004). Recently, researchers have shown an interest in conceptualizing these reasons using SDT (see Vansteenkiste & Mouratidis, 2016). Vansteenkiste, Mouratidis, and Lens (2010) were the first to publish empirical work relying on such a conceptualization. Soccer



players first reported their performance goals (e.g., “It is my goal to perform better than my direct opponent”); then, they reported the autonomous and controlled reasons connected to their performance goals (e.g., “[It is my goal to perform better than my direct opponent] because this goal is a challenge to me,” pp. 223–230). The relations between performance goals and beneficial experiential outcomes were found to drop to nonsignificance (e.g., for positive emotion) or considerably (e.g., for subjective vitality) when controlling for the positive influence of the autonomous reasons connected to performance goals (for comparable results in educational settings, see Gillet, Lafrenière, Vallerand, Huart, & Fouquereau, 2014; Vansteenkiste, Smeets, et al., 2010).

Gillet, Lafrenière, Huyghebaert, and Fouquereau (2015) used this same approach to study the SDT-derived reasons connected to mastery goals. Workers first reported their mastery goals, and then they reported the autonomous and controlled reasons connected to their mastery goals (e.g., “[My goal is to improve] because of the fun and enjoyment that it provides me,” p. 862). The relations between mastery goals and beneficial experiential (e.g., positive emotion) and self-regulated learning (e.g., engagement) outcomes dropped to nonsignificance when controlling for the positive influence of the autonomous reasons connected to mastery goals (see also Gaudreau & Braaten, 2016; for related research with dominant achievement goals, see Michou, Vansteenkiste, Mouratidis, & Lens, 2014; Ozdemir Oz, Lane, & Michou, 2015; Vansteenkiste, Mouratidis, van Riet, & Lens, 2014).

In interpreting these results, researchers commonly state that their methodology has enabled them to detach reasons from goals, and that the autonomous reasons connected to the achievement goals are stronger (Gillet et al., 2015), more robust (Vansteenkiste, Mouratidis, et al., 2010), and more important (Deci & Ryan, 2016) predictors of beneficial outcomes than the achievement goals per se. We do not agree with these interpretations (see also Vansteenkiste, Mouratidis, et al., 2014, for a more nuanced view). We believe that the reason-based variable focused on in the extant work is best represented as an achievement goal complex. An achievement goal complex is a composite motivational construct, comprised of an achievement goal combined with information regarding the reason for pursuing the goal (Elliot & Thrash, 2001). The structural form of an achievement goal complex is “ACHIEVEMENT GOAL because REASON,” which is the typical form of the reason-based variables used in the aforementioned research, for example “MY GOAL IS TO IMPROVE because OF THE FUN AND ENJOYMENT THAT IT PROVIDES ME”.

The consequence of such a reinterpretation is twofold. First, in the approach used to date, autonomous and controlled reasons have only been operationalized with reference to the specific, focal achievement goal; there has been no assessment of reasons in and of themselves, separate from the focal achievement goal. Thus, from our perspective, the results of the existing research actually indicate that autonomous achievement goal complexes eliminate or reduce the influence of achievement goals per se, *not* that autonomous reasons in and of themselves eliminate or reduce the influence of achievement goals per se. Second, it is important to bear in mind that in the approach used to date there is redundancy in the measurement of achievement goals: The achievement goal is assessed multiple times, both alone as a focal goal and in the reason-based variable that connects the goal with reasons (see Senko & Tropiano, 2016, for a related point). Thus, it should not

be surprising that autonomous achievement goal complexes eliminate or reduce the influence of achievement goals per se, because the two variables have overlapping content. In the following, we seek to clarify and extend the existing research by proposing a systematic approach to studying achievement goals, reasons for goal pursuit, and specific achievement goal complexes.

### A Systematic Approach to Studying Goals, Reasons, and Goal Complexes

Goal complexes are multicomponent constructs. In studying them, it is important to carefully distinguish between their component parts and to design assessments accordingly. A first component is the focal goal that represents an aim per se without any accompanying reason. In measurement, it is critical to use a “pure goal” assessment uncontaminated by reason content (e.g., for mastery goals: “My goal is to learn;” see Elliot & Murayama, 2008, on this contamination issue). A second component is the focal reason that represents a more general form of motivation without any specific aim. In measurement, it is critical to also use a “pure reason” assessment uncontaminated by specific goal content (e.g., for autonomous reasons: “I pursue goals because I find them challenging”).<sup>1</sup> Combining the pure goal with the pure reason creates a third construct, the integrated goal complex. It represents an instrumental relation between the goal and the reason: The goal serves the reason and the reason provides the impetus for goal adoption and pursuit. In measurement, this functional relation is explicitly expressed (e.g., for the autonomous mastery goal complex: “My goal is to learn because I find this a highly challenging goal”).<sup>2</sup>

Once these three constructs—goal, reason, and goal complex—are separately assessed, they may be used in three sets of analyses. First, goals and reasons may be tested *separately* to determine their

<sup>1</sup> In the literature, SDT-derived reason assessments are often tied to a generic goal-directed behavior (e.g., “I work because it is fun;” Gagné & Deci, 2005, p. 334). However, goal complex assessments are not tied to a behavior, but to a particular goal (e.g., “In my work, my goal is to learn because I find it fun;” see Vansteenkiste, Lens, et al., 2014). When studying goal complexes, as distinct from other motivational complexes (see Murray, 1938), it is critical to operationalize reasons, goals, and goal complexes in a symmetrical manner: Each motivational construct should be measured with respect to the same reference component. Specifically, in order to isolate the influence of reasons from the influence of goals and goal complexes, SDT-derived reason assessments need to be stripped of behavioral elements and tied to goal regulation in general (e.g., “In my work, I pursue goals because I find them fun;” for such an operationalization, see Sheldon & Elliot, 1998).

<sup>2</sup> In past research, an achievement goal complex was sometimes operationalized as the product term between an achievement goal and a reason variable (e.g., Gaudreau, 2012; for experimental work, see Benita, Roth, & Deci, 2014; Spray, John Wang, Biddle, & Chatzisarantis, 2006). In our approach, however, the product term between the “pure mastery goal” variable and the “pure autonomous reason” variable would *not* correspond to an autonomous mastery goal complex. “Pure mastery goals” may be energized by reasons other than autonomous reasons (e.g., controlled reasons), whereas “pure autonomous reasons” may be directed by goals other than mastery goals (e.g., performance goals), therefore the interaction between mastery goals and autonomous reasons does not necessarily represent an autonomous mastery goal complex. In other words, high mastery goals and high autonomous reasons do not always indicate a high autonomous mastery goal complex, and a third composite variable is needed to capture the extent to which these goals and reasons combine to form a single, inseparable, and additional achievement goal complex variable.



individual links to outcomes. Second, goals and reasons may be tested *simultaneously* to determine their unique links to outcomes. Third, goal complexes may be tested together with goals and reasons to determine the incremental contribution of goal complexes to outcomes, as well as the contribution of goals per se and reasons per se. In the following, we apply this approach to the central constructs studied in our research herein: mastery goals, autonomous reasons, and autonomous mastery goal complexes.

### Testing Mastery Goals and Autonomous Reasons as Separate Predictors

As reviewed earlier, mastery goals and autonomous reasons have been shown to similarly predict beneficial achievement-relevant outcomes. We expected to find the same predictive patterns for mastery goals and autonomous reasons as that found in prior work.

**Hypothesis 1:** Mastery goals (H1a) and autonomous reasons (H1b) are positive predictors of beneficial experiential and self-regulated learning outcomes.

### Testing Mastery Goals and Autonomous Reasons as Simultaneous Predictors

Mastery goals and autonomous reasons are both distinct and overlapping constructs. They are conceptually distinct in that they have unique properties, operate at different levels of specificity, and have different functions. Mastery goals are concrete cognitive representations of future competence-relevant possibilities that proximally direct individuals' behavior (Elliot & Fryer, 2008). Autonomous reasons are general need-based internal forces that provide energy for action (Deci & Ryan, 2008). Furthermore, principal component factor analysis has revealed that mastery goal and autonomous reason items loaded on different factors (Dysvik & Kuvaas, 2010). Given their conceptual and empirical distinctiveness, we expected mastery goals and autonomous reasons to explain independent variance in the beneficial experiential and self-regulated learning outcomes to which they are (separately) linked.

**Hypothesis 2:** Mastery goals (H2a) and autonomous reasons (H2b) explain independent variance in beneficial experiential and self-regulated learning outcomes.

Although they are conceptually and empirically distinct, mastery goals and autonomous reasons are also overlapping constructs. Mastery goals are sometimes described as intrinsic goals (Pintrich & Garcia, 1991) and emerge from autonomy-supportive contexts (Diseth & Samdal, 2014); autonomous reasons are viewed as facilitating the expression of one's agentic tendency to learn (Ryan & Powelson, 1991) and emerge from mastery-focused climates (Standage et al., 2005). Furthermore, a positive correlation is commonly observed between mastery goals and autonomous reasons (e.g., Katz, Assor, & Kanat-Maymon, 2008). Given this conceptual and empirical overlap, the predictive utility of mastery goals should be diminished when partialing out the variance explained by autonomous reasons—this is consistent with the position articulated in the extant research on SDT-derived reasons and achievement goals, but has not yet been tested. Conversely, the predictive utility of autonomous reasons should also be diminished

when partialing out the variance explained by mastery goals—this also has not been tested in the extant research.

**Hypotheses 3:** The predictive strength of mastery goals is diminished when controlling for autonomous reasons (H3a), and the predictive strength of autonomous reasons is diminished when controlling for mastery goals (H3b).

### Testing Autonomous Mastery Goal Complexes Together With Goals and Reasons

According to gestalt principles, a goal complex should be more than the mere sum of a goal and a reason (Lewin, 1951). That is, autonomous reasons combined with a mastery goal should do more than just add an exogenous reason element to the goal, they should alter the functional significance of the goal and the experience of goal regulation (Deci & Ryan, 1985; Elliot, 2006). Both mastery goals and autonomous reasons are commonly portrayed as optimal forms of motivation (Kaplan & Maehr, 2007; Sheldon, 2004), and it is likely that their integration in the form of an achievement goal complex would be particularly beneficial for achievement-relevant outcomes. Autonomous reasons may enhance mastery goal persistence and attainment via challenge appraisals (Ntoumanis et al., 2014), and mastery goals may help maintain a focus on the positive value of the task and facilitate interest-based engagement (Huang, 2011; Senko & Miles, 2008). In other words, autonomous reasons are assumed to predict goal success (i.e., effective goal regulation), and when *specifically* combined with mastery goals, goal success is assumed to further lead to beneficial experiential and self-regulated learning outcomes (i.e., effective behavior regulation). This would be consistent with the findings observed in the extant research on SDT-derived reasons and achievement goals, although in that work autonomous reasons in and of themselves were not accounted for.

**Hypotheses 4:** The autonomous mastery goal complex explains incremental variance in beneficial experiential and self-regulated learning outcomes.

As noted above, there is measurement redundancy when achievement goal complexes and their component parts are assessed. As such, the predictive utility of mastery goals should be diminished when examining the autonomous mastery goal complex—this is how we interpret the findings in the extant research on SDT-derived reasons and achievement goals. Likewise, given the measurement redundancy with regard to autonomous reasons, the predictive utility of autonomous reasons should be diminished when examining the autonomous mastery goal complex—this has not been considered in the extant research.

**Hypotheses 5:** The predictive strength of mastery goals (H5a) and autonomous reasons (H5b) is diminished when controlling for the autonomous mastery goal complex.

### Overview of the Studies

We designed four studies to disentangle the influence of achievement goals (especially mastery goals), reasons (especially autonomous reasons), and achievement goal complexes (especially the autonomous mastery goal complex) on the most commonly



investigated beneficial experiential and self-regulated learning outcomes. In Study 1, we tested Hypotheses 1a–1b, 2a–2b, and 3a–3b (detaching goals from reasons); in Studies 2 to 4, we additionally tested Hypotheses 4 and 5a–5b (detaching goal complexes from goals and reasons). In Studies 1 and 2, we assessed beneficial experiential outcomes (i.e., interest, satisfaction, positive emotion); in Studies 3 and 4, we assessed beneficial self-regulated learning outcomes (i.e., deep learning, help-seeking, challenging tasks, persistence). In Studies 1 to 3, we focused solely on the goal variable of central interest, namely mastery goals; in Study 4, we extended the hypotheses to performance goals and performance goal-relevant outcomes. Studies 1 to 3 were conducted in a work setting; Study 4 was conducted in an educational setting. In each study we also assessed controlled reasons (and associated controlled achievement goal complexes). Given that our research focused on beneficial outcomes and that controlled reasons and controlled goal complexes are more likely to be predictors of detrimental outcomes, no predictions were made for these variables. However, as in prior research, these variables were entered as covariates (e.g., Gillet et al., 2015) and the influence of controlled achievement goal complexes will be addressed in the General Discussion section.

Table 1 provides a summary and guide for the research; it states each hypothesis, its rationale, its operationalized predictor(s), and the studies and outcomes to which it relates. In all studies, sample sizes were determined a priori, and all manipulations, data exclusions, and measures analyzed are reported. Questionnaires, raw data, and syntax files for the four studies are available through FigShare (<https://figshare.com/s/18543835e916a359b33e>).

### Study 1. Mastery Goals, Reasons, and Experiential Outcomes

Study 1 was designed to test mastery goals and SDT-derived reasons as predictors of three experiential outcomes. Participants reported their work-based mastery goals, and their autonomous and controlled reasons for goal pursuit. Participants also reported their job interest, satisfaction, and positive emotion; we assessed these variables with measures used in prior work in this area (Gillet et al., 2015, 2014; Ozdemir Oz et al., 2015).

### Method

**Participants.** Amazon Mechanical Turk (MTurk) was used as the crowdsourcing platform for data collection. MTurk workers are more demographically diverse than standard Internet samples and American undergraduate samples (Buhrmester, Kwang, & Gosling, 2011). An a priori power analysis revealed that 395 participants were needed to detect small-sized effects ( $f^2 = .02$ ) in a multiple linear regression model with power of .80. We oversampled to make sure that we exceeded our target sample size after excluding missing data. To participate, MTurk workers had to currently have a job. A total of 467 participants completed the questionnaire; seven were excluded a priori due to missing data on the outcome variables. The final sample consisted of 460 U.S. residents, 278 men and 181 women (one not reported), with a mean age of 32.18 ( $SD = 9.04$ ), and having held their job for 6.03 years ( $SD = 5.70$ ). Individuals received 0.20 USD for participating.<sup>3</sup>

**Procedure.** Participants stated their current job and reported their work-based mastery goals and reasons for goal pursuit. The goal and reason variables were counterbalanced: 249 participants completed the reason items first, 211 completed the goal items first. Then, job interest, satisfaction, and positive emotion were assessed.

**Measures.** Table 2 presents the descriptive statistics and correlation matrix. Participants responded using a 1 = *not at all*, 4 = *somewhat*, 7 = *completely* scale.

**Mastery goals.** Elliot and Murayama's (2008) Achievement Goal Questionnaire—Revised (AGQ-R) was adapted to assess work-based mastery goals. The three items were presented as “descriptions of how [one] might pursue goals at [his/her] job” (e.g., “In my job, my goal is to learn as much as possible”).

**Autonomous and controlled reasons for goal pursuit.** Michou et al. (2014) measure was adapted to assess work-based autonomous and controlled reasons for goal pursuit. To disentangle the goal component from the reason component, we adjusted these items so that they did *not* refer to a specific achievement goal. The items were presented as “explanations for why [one] might pursue goals at [his/her] job.” Two items assessed autonomous reasons (e.g., “In my job, I pursue goals because I find them highly stimulating and challenging”) and four items assessed controlled reasons (e.g., “In my job, I pursue goals because others will reward me only if I achieve these goals”).

**Job interest.** Ryan's (1982) six-item Intrinsic Motivation Inventory was adapted to assess job interest (e.g., “I would describe my work as very interesting”).

**Job satisfaction.** Diener, Emmons, Larsen, and Griffin's (1985) five-item Satisfaction with Life Scale was adapted to assess job satisfaction (e.g., “I am satisfied with my work”).

**Job positive emotion.** Watson, Clark, and Tellegen's (1988) Positive and Negative Affect Schedule was adapted to assess job positive emotion. Participants were asked to indicate the extent they feel 10 positive emotions in their work (e.g., “excited,” “proud”).

### Results

**Overview.** We used sequential linear regression for our analyses. For each outcome variable, three models were built. First, in the “goal-only” model, only mastery goals were included as a predictor (Model 1 in Table 3). Second, in the “reason-only” model, only autonomous and controlled reasons were included as predictors (Model 2 in Table 3). Third, in the “goal-and-reason” model, mastery goals *and* autonomous and controlled reasons were included as predictors (Model 3 in Table 3). This enabled us to estimate the independent contribution of the two focal variables—mastery goals and autonomous reasons—as well as the reduction of their predictive strength when partialing out the variance accounted for by the other variable.

**Preliminary analysis.** We conducted a preliminary analysis to examine potential covariates: sex (“1” = male, “2” = female, for all studies), age, and seniority. In addition, we tested the

<sup>3</sup> For this and the subsequent studies, the payment was way well above the reservation wage of \$1.38 per hour (i.e., the minimum wage a worker is willing to accept to complete a task; Horton & Chilton, 2010). Payment level has been found not to affect data quality (Buhrmester et al., 2011).

Table 1  
Summary of the Hypotheses, Their Rationale, Their Operationalized Predictors, and the Studies and Outcomes to Which They Relate

Hypotheses	Rationale	Predictors and "operationalization"	Studies: Types of outcome
H1a. Mastery goals are a positive predictor of beneficial outcomes	Replication of prior research	Mastery goals alone "My goal is to learn"	S1-2: Experiential S3-4: Self-regulated learning S4: Extended to performance goals
H1b. Autonomous reasons are a positive predictor of beneficial outcomes	Replication of prior research	Autonomous reasons alone "I pursue goals because I find them challenging"	S1-2: Experiential S3-4: Self-regulated learning
H2a-b. Mastery goals (H2a) and autonomous reasons (H2b) explain independent variance in beneficial outcomes	Mastery goals and autonomous reasons differ	Mastery goals <i>plus</i> autonomous reasons	S1-2: Experiential S3-4: Self-regulated learning S4: Extended to performance goals
H3a-b. The influence of mastery goals is diminished when controlling for autonomous reasons (H3a), and vice versa (H3b)	Mastery goals and autonomous reasons overlap		S1-2: Experiential S3-4: Self-regulated learning
H4. The autonomous mastery goal complex explains incremental variance in beneficial outcomes	The autonomous mastery goal complex is more than the mere sum of goal and reason	Mastery goals <i>plus</i> autonomous reasons <i>plus</i> autonomous mastery goal complex "My goal is to learn because I find this a highly challenging goal"	S2: Experiential S3-4: Self-regulated learning S4: Extended to performance goals
H5a-b. The influence of mastery goals (H5a) and autonomous reasons (H5b) is diminished when controlling for the autonomous mastery goal complex	Measurement redundancy		S2: Experiential S3-4: Self-regulated learning S4: Extended to performance goals

interactions between order ("1" = reasons first, "2" = goals first, for all studies) and our predictor variables (i.e., mastery goals and autonomous and controlled reasons; see Yzerbyt, Muller, & Judd, 2004). None of the covariates attained significance ( $ps \geq .088$ ), and neither order main nor interactive effects were observed ( $ps \geq .152$ ). Hence these terms were not considered further (including them did not change the pattern of results).

**Main analyses.** For this and all subsequent studies, our report of the results is hypothesis driven. Nontheoretically relevant findings are not reported in the narrative, but are included in Table 3 (which presents the full set of results). Effect size estimates are also included in the tables. These estimates are partial eta squared ( $\eta_p^2$ ), that is, the proportion of variance *uniquely* explained by a

predictor (i.e., while partialing out the effect of the other predictors).

**"Goal-only" model.** In line with Hypothesis 1a, mastery goals were a positive predictor of interest,  $B = 0.62$  [0.53, 0.71],  $p < .001$ , satisfaction,  $B = 0.52$  [0.42, 0.63],  $p < .001$ , and positive emotion,  $B = 0.57$  [0.49, 0.67],  $p < .001$  (numbers in brackets represents 95% confidence intervals).

**"Reason-only" model.** In line with Hypothesis 1b, autonomous reasons were a positive predictor of interest,  $B = 0.66$  [0.59, 0.73],  $p < .001$ , satisfaction,  $B = 0.62$  [0.54, 0.70],  $p < .001$ , and positive emotion,  $B = 0.58$  [0.51, 0.64],  $p < .001$ .

**"Goal-and-reason" model.** In line with Hypothesis 2a, mastery goals remained a positive predictor of interest,  $B = 0.26$  [0.16, 0.36],  $p < .001$ , and positive emotion,  $B = 0.20$  [0.10, 0.30],  $p < .001$ .

Table 2  
Studies 1 and 2: Descriptive Statistics and Correlation Matrix for the Main Variables

	Descriptive statistics (Study 1/Study 2)			Correlation matrix (Study 1 below the diagonal, Study 2 above the diagonal).							
	$\alpha$	$M$	$SD$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mastery goals (1)	.87/.84	5.84/5.85	1.13/1.05	—	.65***	.32***	.73***	.47***	.58***	.49***	.58***
Autonomous reasons (2)	.86/.80	5.33/5.51	1.38/1.21	.60***	—	.28***	.81***	.37***	.64***	.67***	.67***
Controlled reasons (3)	.65/.70	4.85/4.96	1.14/1.19	.28***	.26***	—	.30***	.83***	.07	.29***	.28***
Autonomous mastery goal complex (4)	n/a/.91	n/a/5.48	n/a/1.11	n/a	n/a	n/a	—	.42***	.62***	.60***	.66***
Controlled mastery goal complex (5)	n/a/.91	n/a/5.05	n/a/1.13	n/a	n/a	n/a	n/a	—	.21***	.36***	.38***
Job interest (6)	.88/.84	5.02/5.07	1.31/1.22	.54***	.68***	.11*	n/a	n/a	—	.71***	.68***
Job satisfaction (7)	.91/.89	4.91/5.12	1.43/1.33	.41***	.61***	.19***	n/a	n/a	.74***	—	.71***
Job positive emotion (8)	.94/.94	5.32/5.54	1.26/1.16	.52***	.66***	.26***	n/a	n/a	.78***	.76***	—

Note. n/a = applicable (i.e., the variable was not measured in the study).  
\*  $p < .05$ . \*\*\*  $p < .001$ .



Table 3  
Studies 1 and 2: Coefficient Estimates and Effect Sizes for the Models Testing the Influence of Mastery Goals Alone (Model 1; “Goal-Only” Model), Autonomous and Controlled Reasons Alone (Model 2; “Reason-Only” Model), Mastery Goals and Reasons (Model 3; “Goal-and-Reason” Model), and Mastery Goals, Reasons, and Mastery Goal Complexes (for Study 2: Model 4; “Goal Complex” Model)

	Job interest						Job satisfaction						Job positive emotion					
	Model 1		Model 2		Model 3		Model 1		Model 2		Model 3		Model 1		Model 2		Model 3	
	B	$\eta_p^2$	B	$\eta_p^2$	B	$\eta_p^2$	B	$\eta_p^2$	B	$\eta_p^2$	B	$\eta_p^2$	B	$\eta_p^2$	B	$\eta_p^2$	B	$\eta_p^2$
Study 1																		
Intercept	1.37	—	1.86	—	1.15	—	1.84	—	1.38	—	1.13	—	1.94	—	1.73	—	1.18	—
Mastery goals (MAp)	.62	***	.29	>	.26	***	.52	***	.17	>	.09	—	.58	***	.27	>	.20	***
Autonomous reasons			.66	***	.46	>			.62	***	.35	=			.58	***	.40	>
Controlled reasons			-.07	†	—	—			.05	—	.03	—			.11	**	.01	.08
Study 2																		
Intercept	1.13	—	1.92	—	1.08	—	1.01	—	.64	—	.46	—	.39	—	1.67	—	1.07	—
Mastery goals (MAp)	.67	***	.33	>	.37	***	.28	***	.70	***	.08	—	.00	—	.65	***	.27	***
Autonomous reasons			.68	***	.42	>	.39	***	.12	***	.66	***	.56	***	.61	***	.47	***
Controlled reasons			-.12	**	.02	—	-.24	***	.03	—	.11	*	.05	—	.10	**	.07	†
Autonomous MAp complex							.18	*					.18	*			.24	***
Controlled MAp complex							.09	—					.08	—			.11	—

Note. Variables are not centered. “>” means that the predictive strength of mastery goals in Model 1 is significantly greater than the predictive strength of mastery goals in Model 3 (i.e., there is a significant reduction from Model 1 to Model 3); “=” means that the difference is not significant. This is the case for the other model comparisons (i.e., Model 2 vs. 3, and Model 3 vs. 4) and variable (i.e., autonomous reasons) as well.

†  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

.001; contrary to the hypothesis, mastery goals no longer predicted satisfaction,  $B = 0.09 [-0.02, 0.21]$ ,  $p = .117$ . In line with Hypothesis 2b, autonomous reasons remained a positive predictor of interest,  $B = 0.54 [0.46, 0.62]$ ,  $p < .001$ , satisfaction,  $B = 0.58 [0.48, 0.67]$ ,  $p < .001$ , and positive emotion,  $B = 0.49 [0.41, 0.56]$ ,  $p < .001$ .

In this and the subsequent studies, we used the Monte Carlo method (with 50,000 simulations) to estimate the confidence intervals for reduction of the predictive strength of mastery goals when controlling for autonomous reasons, and vice versa (MacKinnon, Lockwood, & Williams, 2004). In addition, percentage reductions in the effect and Sobel tests are reported in parentheses ( $Z$  tests and  $p$  values). In line with Hypothesis 3a, the reduction of the relations between mastery goals and interest,  $B = 0.38 [0.31, 0.45]$  (59% reduction), satisfaction,  $B = 0.40 [0.32, 0.42]$  (81%), and positive emotion,  $B = 0.34 [0.27, 0.41]$  (63%), due to the inclusion of autonomous reasons were significant ( $Z_s \geq 9.30$ ,  $p_s < .001$ ). In line with Hypothesis 3b, the reduction of the relations between autonomous reasons and interest,  $B = 0.12 [0.07, 0.17]$  (18%), and positive emotion,  $B = 0.09 [0.05, 0.14]$  (16%), due to the inclusion of mastery goals were significant ( $Z_s \geq 3.96$ ,  $p_s < .001$ ); contrary to the hypothesis, the reduction of the relation between autonomous reasons and satisfaction,  $B = 0.04 [-0.01, 0.10]$  (7%), was not significant ( $Z = 1.56$ ,  $p = .118$ ).

## Discussion

Mastery goals (Hypothesis 1a) and autonomous reasons (Hypothesis 1b) accounted for variance in interest, satisfaction, and positive emotion *when tested separately*. More importantly, mastery goals (Hypothesis 2a) and autonomous reasons (Hypothesis 2b) each explained independent variance in interest and positive emotion *when tested simultaneously*. Moreover, the predictive strength of mastery goals (Hypothesis 3a) and autonomous reasons (Hypothesis 3b) for interest and positive emotion were diminished *when taking the other into account*. This suggests that neither construct “captured” all of the variance explained by the other: Mastery goals and autonomous reasons shared predictive utility with regard to these outcomes, but their overlap was not so substantial as to conclude that one eliminates the influence of the other. For satisfaction, however, Hypothesis 2a and 3b were not supported. Mastery goals no longer explained a significant portion of variance in satisfaction when autonomous reasons were controlled, and controlling for mastery goals did not significantly diminish the influence of autonomous reasons. This suggests that for at least some outcomes, the influence of reasons may indeed outweigh the influence of goals.

One important issue that Study 1 left unaddressed is the autonomous mastery goal complex. Prior goal complex research has shown (from our perspective) that controlling for the autonomous mastery goal complex leads to a decrease in the predictive strength of mastery goals; however, it has not tested for a parallel decrease in the predictive strength of autonomous reasons. In Study 2, we unambiguously separate achievement goals, reasons, and achievement goal complexes in order to test whether the autonomous mastery goal complex explains incremental variance in interest, satisfaction, and positive emotion, and whether it diminishes the predictive strength of both mastery goals and autonomous reasons.

## Study 2. Mastery Goals, Reasons, Goal Complexes, and Experiential Outcomes

Study 2 was designed to test mastery goals, SDT-derived reasons, and achievement goal complexes as predictors of the same experiential outcomes used in Study 1. Participants reported their work-based mastery goals, their autonomous and controlled reasons for goal pursuit, and their autonomous and controlled mastery goal complexes. Participants also reported their job interest, satisfaction, and positive emotion.

## Method

**Participants.** The target sample size was the same as in Study 1. To participate, MTurk workers had to currently have a job and not have participated in Study 1. A total of 407 participants completed the questionnaire; one was excluded a priori due to missing data on the outcome variables. The final sample consisted of 406 U.S. residents, 236 men and 170 women, with a mean age of 33.18 ( $SD = 10.07$ ), and having held their job for 6.36 years ( $SD = 5.87$ ). Individuals received 0.20 USD for participating.

**Procedure.** Participants stated their current job and reported their work-based mastery goals, reasons, and goal complexes. As in Study 1, the goal and reason variables were counterbalanced: 206 participants completed the reason items first, 200 completed the goal items first. Then, job interest, satisfaction, and positive emotion were assessed.

**Measures.** Table 2 presents the descriptive statistics and correlation matrix. Participants responded using a 1 = *not at all*, 4 = *somewhat*, 7 = *completely* scale.

**Mastery goals.** The same measure used in the prior study was used in this study.

**Autonomous and controlled reasons for goal pursuit.** The same measure used in the prior study was used in this study.

**Autonomous and controlled mastery goal complexes.** Each of the three items measuring mastery goals were combined with each of the six items measuring autonomous and controlled reasons to assess work-based autonomous and controlled mastery goal complexes. The statements thus produced were presented as “descriptions of how you might pursue goals at your job, together with explanations for why you might pursue them.” Six items (3 goal items  $\times$  2 reason items) assessed the autonomous mastery goal complex (e.g., “In my job, my goal is to learn as much as possible because I find this a highly stimulating and challenging goal”), and 12 items (3 goal items  $\times$  4 reason items) assessed the controlled mastery goal complex (e.g., “In my job, my goal is to learn as much as possible because others will reward me only if I achieve this goal”).

**Job interest, satisfaction, and positive emotion.** Job interest, satisfaction, and positive emotion were assessed using the same measures used in Study 1.

## Results

**Overview.** We used the same analytical strategy as in Study 1, albeit with a fourth step added to test the “goal complex” model. In this model, mastery goals, autonomous and controlled reasons, and autonomous and controlled mastery goal complexes were included as predictors (Model 4 in Table 3). This enabled us to



estimate the incremental contribution of the autonomous mastery goal complex, as well as the reduction of the predictive strength of mastery goals and autonomous reasons when controlling for this goal complex.<sup>4</sup>

**Preliminary analysis.** As in Study 1, we conducted a preliminary analysis to examine potential covariates (sex, age, seniority) and order effects. None of the covariates attained significance ( $ps \geq .061$ ), excepting a positive association between seniority and interest,  $B = 0.02$  [0, 0.04],  $p = .025$ . Although no order main effects were observed ( $ps \geq .634$ ), order interacted with mastery goals in predicting interest,  $B = -0.26$  [-0.49, -0.04],  $p = .021$ , and with autonomous reasons in predicting interest,  $B = 0.23$  [0.03, 0.42],  $p = .021$ , and positive emotion,  $B = 0.19$  [0.01, 0.37],  $p = .042$ . As including these terms was neither theoretically relevant nor changed the pattern of results, they were not considered further.

**Main analyses.** Table 3 presents the full set of results.

**“Goal-only” model.** In line with Hypothesis 1a, mastery goals were a positive predictor of interest,  $B = 0.67$  [0.58, 0.77],  $p < .001$ ; satisfaction,  $B = 0.62$  [0.51, 0.73],  $p < .001$ ; and positive emotion,  $B = 0.65$  [0.56, 0.73],  $p < .001$ .

**“Reason-only” model.** In line with Hypothesis 1b, autonomous reasons were a positive predictor of interest,  $B = 0.68$  [0.60, 0.76],  $p < .001$ ; satisfaction,  $B = 0.70$  [0.62, 0.79],  $p < .001$ ; and positive emotion,  $B = 0.61$  [0.54, 0.68],  $p < .001$ .

**“Goal-and-reason” model.** In line with Hypothesis 2a, mastery goals remained a positive predictor of interest,  $B = 0.37$  [0.26, 0.48],  $p < .001$ , and positive emotion,  $B = 0.27$  [0.16, 0.37],  $p < .001$ ; contrary to the hypothesis, mastery goals no longer predicted satisfaction,  $B = 0.08$  [-0.04, 0.20],  $p = .195$ . In line with Hypothesis 2b, autonomous reasons remained a positive predictor of interest,  $B = 0.49$  [0.39, 0.58],  $p < .001$ ; satisfaction,  $B = 0.66$  [0.55, 0.77],  $p < .001$ ; and positive emotion,  $B = 0.47$  [0.38, 0.56],  $p < .001$ .

In line with hypothesis 3a, the Monte Carlo method revealed that the reduction of the relations between mastery goals and interest,  $B = 0.35$  [0.28, 0.44] (49% reduction), satisfaction,  $B = 0.48$  [0.39, 0.58] (86%), and positive emotion,  $B = 0.34$  [0.27, 0.42] (56%), due to the inclusion of autonomous reasons were significant ( $Zs \geq 8.54$ ,  $ps < .001$ ). In line with Hypothesis 3b, the reduction of the relations between autonomous reasons and both interest,  $B = 0.19$  [0.13, 0.26] (29%), and positive emotion,  $B = 0.14$  [0.08, 0.20] (23%), due to the inclusion of mastery goals were significant ( $Zs \geq 4.75$ ,  $ps < .001$ ); contrary to the hypothesis, the reduction in the relation between autonomous reasons and satisfaction,  $B = 0.04$  [-0.02, 0.11] (6%), was not significant ( $Z = 1.29$ ,  $p = .196$ ).

**“Goal complex” model.** In line with Hypothesis 4, the autonomous mastery goal complex was a positive predictor of interest,  $B = 0.18$  [0.03, 0.33],  $p = .015$ ; satisfaction,  $B = 0.18$  [0.02, 0.34],  $p = .031$ ; and positive emotion,  $B = 0.24$  [0.10, 0.38],  $p < .001$ .

Again, we used the Monte Carlo method to estimate the reduction of the predictive strength of mastery goals and autonomous reasons when controlling for the autonomous mastery goal complex. In line with Hypothesis 5a, the reduction of the relations between mastery goals and both interest  $B = 0.06$  [0.01, 0.11] (18%), and positive emotion  $B = 0.08$  [0.03, 0.13] (34%), due to the inclusion of the autonomous mastery goal complex were sig-

nificant ( $Zs \geq 2.34$ ,  $ps \leq .019$ ; mastery goals remained a significant predictor in both instances,  $ps \leq .01$ ). The analysis was not conducted for satisfaction, given the null relation for mastery goals in the “goal-and-reason” model. In line with Hypothesis 5b, the reduction of the relations between autonomous reasons and interest,  $B = 0.10$  [0.02, 0.17] (20%), satisfaction,  $B = 0.09$  [0.01, 0.18] (14%), and positive emotion,  $B = 0.13$  [0.05, 0.20] (27%), due to the inclusion of the autonomous mastery goal complex were significant ( $Zs \geq 2.14$ ,  $ps \leq .032$ ; autonomous reasons remained a significant predictor in all instances,  $ps < .001$ ).

## Discussion

Replicating Study 1’s findings, mastery goals and autonomous reasons accounted for variance in interest, satisfaction, and positive emotion when tested separately, and also explained independent variance in interest and positive emotion when controlling for the other variable (with the predictive strength of each being diminished). This suggests that mastery goals and autonomous reasons overlap without canceling one another. However, as in Study 1, satisfaction was more robustly predicted by autonomous reasons than by mastery goals.

Extending Study 1’s findings, the autonomous mastery goal complex explained incremental variance in interest, satisfaction, and positive emotion (Hypothesis 4). Thus, mastery goals and autonomous reasons not only have an independent influence on adaptive outcomes, they fuse together in the form of a goal complex that has additional predictive benefits. Moreover, the predictive strength of mastery goals (Hypothesis 5a) and autonomous reasons (Hypothesis 5b) were diminished when controlling for the autonomous mastery goal complex. In line with Gillet et al. (2015) findings (from our perspective), controlling for the autonomous mastery goal complex diminishes the predictive strength of mastery goals per se; however, it also diminishes the predictive strength of autonomous reasons per se.

The effect sizes for mastery goals were descriptively smaller than those for autonomous reasons. One possible reason for this is the nature of the outcome variables used in the first two studies. Building on existing research, we used experiential outcomes, which may be particularly sensitive to feelings of task autonomy (Ryan & Deci, 2006). In Study 3, we switched to self-regulated learning outcomes, which may be equally sensitive to mastery goals and autonomous reasons (see Dysvik & Kuvaas, 2013). Specifically, in Study 3 we tested the same set of five hypotheses with the following self-regulated learning outcomes: deep learning, interpersonal help-seeking behavior, and challenging tasks.

## Study 3. Mastery Goals, Reasons, Goal Complexes, and Self-Regulated Learning

Study 3 was designed to test mastery goals, SDT-derived reasons, and achievement goal complexes as predictors of three

<sup>4</sup> Vansteenkiste, Smeets, et al. (2010) noted that variables connecting autonomous or controlled reasons to a given achievement goal could seem odd for a participant not pursuing this achievement goal. Accordingly, we repeated the analyses for the full study, excluding the two participants with an average mastery goal score below 2 (3 in Study 3; 6 in Study 4). The results for the achievement goal complex variables remained essentially the same as those reported in the text (this is the case for all studies).

self-regulated learning outcomes. Participants reported their work-based mastery goals, their autonomous and controlled reasons for goal pursuit, and their autonomous and controlled mastery goal complexes. They also reported their job deep learning, help-seeking, and challenging tasks.

## Method

**Participants.** The target sample size was the same as in the prior studies. To participate, MTurk workers had to currently have a job and not have participated in Studies 1 or 2. A total of 440 participants completed the questionnaire; 11 were excluded a priori due to missing data on the outcome variables. The final sample consisted of 429 U.S. residents, 213 men and 216 women, with a mean age of 34.19 ( $SD = 10.07$ ), and having held their job for 6.23 years ( $SD = 6.64$ ). Individuals received 0.30 USD for participating.

**Procedure.** Participants stated their current job and reported their work-based mastery goals, reasons, and goal complexes. Again, the goal and reason variables were counterbalanced: 211 participants completed the reason items first, 218 completed the goal items first. Then, job deep learning, help-seeking, and challenging tasks were assessed.

**Measures.** Table 4 presents the descriptive statistics and correlation matrix. Participants responded using a 1 = *not at all*, 4 = *somewhat*, 7 = *completely* scale.

**Mastery goals.** The same measure used in prior study was used in this study.

**Autonomous and controlled reasons for goal pursuit.** The same measure used in the prior study was used in this study.

**Autonomous and controlled mastery goal complexes.** The same measure used in the prior study was used in this study.

**Job deep learning.** Kirby, Knapper, Evans, Carty, and Gadula's (2003) 10-item deep subscale from the Approaches to Learning at Work Questionnaire assessed job deep learning (e.g., "I spend a good deal of my spare time learning about things related to my work").

**Job help-seeking.** Holman, Epitropaki, and Fernie's (2001) three-item interpersonal help seeking subscale from the Scale of Learning Strategies in the Workplace assessed job help-seeking (e.g., "I ask others for more information when I need it [at my work]").

**Job challenging tasks.** Preenen, De Pater, Van Vianen, and Keijzer's (2011) six-item Challenging Assignments Scale was

adapted to assess job challenging tasks (e.g., "[In my work I perform tasks] that are challenging").

## Results

**Overview.** We used the same analytical strategy used in Study 2. For each outcome variable, four linear regression models were built (see Models 1 to 4 in Table 5).

**Preliminary analysis.** As in Studies 1 and 2, we conducted a preliminary analysis to examine potential covariates (sex, age, seniority) and order effects. None of the covariates attained significance ( $ps \geq .083$ ), excepting a negative association between age and deep learning,  $B = -0.02 [-0.02, -0.01]$ ,  $p < .001$ , and a positive association between sex and help-seeking,  $B = 0.20 [0.01, 0.38]$ ,  $p < .001$ . An order main effect was observed on help-seeking,  $B = 0.20 [0.01, 0.40]$ ,  $p = .043$ , as well as an interactive effect with autonomous reasons on deep learning,  $B = -0.13 [-0.25, -0.02]$ ,  $p = .022$ . As including these terms was neither theoretically relevant nor changed the pattern of results, they were not considered further.

**Main analyses.** Table 5 presents the full set of results.

**"Goal-only" model.** In line with Hypothesis 1a, mastery goals were a positive predictor of deep learning,  $B = 0.50 [0.43, 0.58]$ ,  $p < .001$ ; help-seeking,  $B = 0.38 [0.30, 0.46]$ ,  $p < .001$ ; and challenging tasks,  $B = 0.50 [0.42, 0.58]$ ,  $p < .001$ .

**"Reason-only" model.** In line with Hypothesis 1b, autonomous reasons were a positive predictor of deep learning,  $B = 0.42 [0.37, 0.47]$ ,  $p < .001$ ; help-seeking,  $B = 0.16 [0.09, 0.22]$ ,  $p < .001$ ; and challenging tasks,  $B = 0.37 [0.32, 0.43]$ ,  $p < .001$ .

**"Goal-and-reason" model.** In line with Hypothesis 2a, mastery goals remained a positive predictor of deep learning,  $B = 0.26 [0.18, 0.34]$ ,  $p < .001$ ; help-seeking,  $B = 0.36 [0.26, 0.46]$ ,  $p < .001$ ; and challenging tasks,  $B = 0.28 [0.19, 0.37]$ ,  $p < .001$ . In line with Hypothesis 2b, autonomous reasons remained a positive predictor of deep learning,  $B = 0.32 [0.26, 0.38]$ ,  $p < .001$ , and challenging tasks,  $B = 0.27 [0.20, 0.33]$ ,  $p < .001$ ; contrary to the hypothesis, these reasons no longer predicted help-seeking  $B = 0.02 [-0.05, 0.09]$ ,  $p = .560$ .

In line with hypothesis 3a, the Monte Carlo method revealed that the reduction of the relations between mastery goals and both deep learning,  $B = 0.23 [0.18, 0.28]$  (46% reduction), and challenging tasks,  $B = 0.19 [0.14, 0.25]$  (41%), due to the inclusion of autonomous reasons were significant ( $Zs \geq 6.82$ ,  $ps < .001$ ); contrary to the hypothesis, the reduction in the relation between

Table 4

Study 3: Descriptive Statistics and Correlation Matrix for the Main Variables

	Descriptive statistics			Correlation matrix							
	$\alpha$	$M$	$SD$	(1)	(2)	(3)	(4)	(5) †	(6)	(7)	(8)
Mastery goals (1)	.88	5.89	1.18	1.00							
Autonomous reasons (2)	.87	5.02	1.56	.54***	1.00						
Controlled reasons (3)	.66	4.67	1.24	.30***	.18***	1.00					
Autonomous mastery goal complex (4)	.91	5.22	1.44	.64***	.82***	.16***	1.00				
Controlled mastery goal complex (5)	.95	4.68	1.23	.32***	.21***	.79***	.24***	1.00			
Job deep learning strategy (6)	.87	4.90	1.08	.55***	.62***	.22***	.70***	.31***	1.00		
Job interpersonal help-seeking (7)	.88	5.91	1.09	.42***	.25***	.16***	.31***	.18***	.28***	1.00	
Job challenging tasks (8)	.85	5.50	1.13	.52***	.54***	.25***	.57***	.28***	.57***	.42***	1.00

\*\*\*  $p < .001$ .



Table 5

Study 3: Coefficient Estimates and Effect Sizes for the Models Testing the Influence of Mastery Goals Alone (Model 1; “Goal-Only” Model), Autonomous and Controlled Reasons Alone (Model 2; “Reason-Only” Model), Mastery Goals and Reasons (Model 3; “Goal-and-Reason” Model), and Mastery goals, reasons, and Mastery Goal Complexes (Model 4; “Goal Complex” Model)

	Job deep learning strategies						Job interpersonal help-seeking						Job challenging tasks					
	Model 1		Model 2		Model 3		Model 4		Model 1		Model 2		Model 3		Model 4		Model 1	
	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$
Intercept	1.91	—	2.33	—	1.53	—	1.35	—	3.64	—	4.63	—	3.54	—	2.57	—	2.10	—
Mastery goals (MAp)	.51***	.31	>	.26***	.09	>	.14**	.03	.38***	.17	.16***	.05	.36***	.11	.33***	.08	.28***	.08
Autonomous reasons			.42***	.37	.32***	.21	.10*	.02					.02		-.03		.27***	.13
Controlled reasons			.10**	.02	.04	—	-.06	—			.10*	.01	.03	—	.01		.09*	.01
Autonomous MAp complex							.34***	.10							.08		.18**	.02
Controlled MAp complex							.16***	.02							.04		.06	

Note. Variables are not centered. “>” means that the predictive strength of mastery goals in Model 1 is significantly greater than the predictive strength of mastery goals in Model 3 (i.e., there is a significant reduction from Model 1 to Model 3); “=” means that the difference is not significant. This is the case for the other model comparisons (i.e., Model 2 vs. 3, and Model 3 vs. 4) and variable (i.e., autonomous reasons) as well.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

mastery goals and help-seeking,  $B = 0.02 [-0.04, 0.07]$  (4%), was not significant ( $Z < 1$ ,  $p = .560$ ). In line with Hypothesis 3b, the reduction of the relations between autonomous reasons and deep learning,  $B = 0.10 [0.07, 0.14]$  (24%); help-seeking,  $B = 0.14 [0.10, 0.18]$  (87%); and challenging tasks,  $B = 0.11 [0.07, 0.14]$  (28%) due to the inclusion of mastery goals were significant ( $Z_s \geq 5.52$ ,  $ps < .001$ ).

“Goal complex” model. In line with Hypothesis 4, the autonomous mastery goal complex was a positive predictor of deep learning,  $B = 0.34 [0.24, 0.43]$ ,  $p < .001$ , and challenging tasks,  $B = 0.18 [0.07, 0.30]$ ,  $p = .001$ ; contrary to the hypothesis, the autonomous mastery goal complex did not predict help-seeking,  $B = 0.08 [-0.04, 0.21]$ ,  $p = .205$ .

In line with Hypothesis 5a, the Monte Carlo method revealed that the reduction of the relations between mastery goals and both deep learning,  $B = 0.11 [0.07, 0.15]$  (45%), and challenging tasks,  $B = 0.06 [0.02, 0.10]$  (23%), due to the inclusion of the autonomous mastery goal complex were significant ( $Z_s \geq 3.01$ ,  $ps \leq .003$ ; mastery goals remained a significant predictor in both instances,  $ps \leq .001$ ). In line with Hypothesis 5b, the reduction of the relations between autonomous reasons and both deep learning,  $B = 0.21 [0.15, 0.27]$  (67%), and challenging tasks,  $B = 0.11 [0.04, 0.18]$  (43%), due to the inclusion of the autonomous mastery goal complex were significant ( $Z_s \geq 3.17$ ,  $ps \leq .002$ ; autonomous reasons remained a significant predictor in both instances,  $ps \leq .011$ ). The analysis was not conducted for help-seeking, given the null relation for the autonomous mastery goal complex.

Discussion

Consistent with Studies 1 and 2, mastery goals and autonomous reasons accounted for variance in deep learning, help-seeking, and challenging tasks when tested separately, and also explained independent variance in deep learning and challenging tasks when tested simultaneously (with the predictive strength of each being diminished). For help-seeking, however, predictions were not supported. Autonomous reasons no longer explained a significant portion of variance in help-seeking when mastery goals were controlled for, and controlling for autonomous reasons did not significantly diminish the influence of mastery goals. Together with the Studies 1 and 2’s findings for satisfaction, this indicates that autonomous reasons may be a more reliable predictor of some variables (satisfaction) and mastery goals a more reliable predictor of others (help-seeking). Rather than concluding that one construct unilaterally reduces the predictive utility of the other, it seems best to view both as important predictors that vary in strength as a function of the outcome in question.

Moreover, consistent with Study 2’s findings, the autonomous mastery goal complex explained additional variance in deep learning and challenging tasks (but not help-seeking), and diminished the predictive strength of mastery goals and autonomous reasons. Thus, again, the autonomous mastery goal complex seems important to consider, and it seems to capture some of the variance explained by mastery goals per se and autonomous reasons per se.

We conducted Study 4 in the academic domain rather than the work domain (see Van Yperen et al., 2014, on the importance of attending to different achievement domains). Study 4 had a threefold aim. First, we sought to test the robustness of Study 3’s findings regarding mastery goals, autonomous reasons, and the autonomous

mastery goal complex as predictors of deep learning and challenging tasks. Second, we sought to extend Studies 1–3’s findings by testing our hypotheses with performance goals. In doing so, we included two outcome variables that performance goals have been shown to positively predict in prior research: surface learning and grade aspiration (Elliot & McGregor, 2001; McGregor & Elliot, 2002). Third, we sought to include an additional outcome variable relevant to mastery goals, performance goals, and autonomous reasons, namely study persistence (Elliot, McGregor, & Gable, 1999; Vallerand et al., 1997). We tested all mastery and performance goal hypotheses in multiple regression models with both goals included, thereby allowing us to determine the influence of each goal while controlling for the influence of the other.

Study 4. Achievement Goals, Reasons, Goal Complexes, and Self-Regulated Learning

Study 4 was designed to test achievement goals, SDT-derived reasons, and achievement goal complexes as predictors of five self-regulated learning outcomes in an academic context. Students reported their academic mastery and performance goals, their autonomous and controlled reasons for goal pursuit, and their autonomous and controlled mastery and performance goal complexes. Participants also reported their deep learning, surface learning, challenging tasks, grade aspiration, and study persistence.

First, all hypotheses were the same for mastery goals, autonomous reasons, and the autonomous mastery goal complex predicting *deep learning* and *challenging tasks*. Second, the hypotheses were extended to performance goals. Performance goals were expected to be a positive predictor of *surface learning* and *grade aspiration* (Hypothesis 1a), even when controlling for autonomous reasons (Hypothesis 2a). Because autonomous reasons are neither compatible nor incompatible with these outcomes (e.g., Donche, Maeyer, Coertjens, Van Daal, & Van Petegem, 2013; Kusrurkar, Ten Cate, Vos, Westers, & Croiset, 2013), Hypotheses 1b, 2b, 3a, and 3b, were not formulated. However, as autonomous reasons may be an ideal motivational foundation from which to efficiently pursue performance goals, the autonomous performance goal complex was expected to explain independent variance in

surface learning and grade aspiration (Hypothesis 4), and to lead to a decrease in the predictive strength of performance goals (Hypothesis 5a). Given the absence of Hypothesis 1b, Hypothesis 5b was not formulated. Third, mastery goals (Hypothesis 1a), performance goals (Hypothesis 1a), and autonomous reasons (Hypothesis 1b) were each expected to be a positive predictor of *study persistence*; accordingly, all remaining hypotheses (Hypotheses 2–5) applied to the relations between the focal predictor variables (mastery goals, performance goals, autonomous reasons, and the autonomous achievement goal complexes) and study persistence.

Method

**Participants.** The target sample size was the same as in the prior studies. The study was administered via the SONA Psychology Research Participation System of a medium-sized U.S. university. A total of 481 participants completed the questionnaire; 24 were excluded a priori due to missing data on the outcome variables. The final sample consisted of 457 students from various study fields, 103 men and 354 women, with a mean age of 20.21 (*SD* = 1.77), 81 of which were freshmen, 135 sophomores, 118 juniors, and 122 seniors (1 “other”). Individuals received 0.5 extra course credit for participating.

**Procedure.** Participants reported their academic achievement goals, reasons, and goal complexes. Again, the goal and reason variables were counterbalanced: 234 participants completed the reason items first, 223 completed the goal items first. Then, deep and surface learning, challenging tasks, grade aspiration, and study persistence were assessed.

**Measures.** Table 6 presents the descriptive statistics and correlation matrix. Participants responded using a 1 = *not at all*, 4 = *somewhat*, 7 = *completely* scale, unless otherwise specified. The items for all predictor variables are provided in the Appendix.

**Mastery and performance goals.** Elliot and Murayama’s (2008) AGQ-R was used to assess mastery and performance goals. To keep the achievement goal complex variables at a reasonable length, we used only two items to assess mastery goals and two

Table 6  
Study 4: Descriptive Statistics and Correlation Matrix for the Main Variables

	Descriptive statistics			Correlation matrix												
	α	<i>M</i>	<i>SD</i>	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Mastery goals (1)	.78	5.40	1.19	1.00												
Performance goals (2)	.79	5.21	1.30	.36***	1.00											
Autonomous reasons (3)	.77	5.15	1.14	.62***	.30***	1.00										
Controlled reasons (4)	.70	4.32	1.17	.10*	.39***	.10*	1.00									
Autonomous mastery goal complex (5)	.88	5.18	1.10	.73***	.30***	.73***	.08†	1.00				§				
Controlled mastery goal complex (6)	.87	4.21	1.17	.13**	.39***	.09†	.85***	.14**	1.00							
Autonomous performance goal complex (7)	.88	4.74	1.31	.29***	.60***	.36***	.33***	.42***	.39***	1.00						
Controlled performance goal complex (8)	.90	4.22	1.27	-.01	.49***	.02	.72***	.02	.79***	.53***	1.00					
Deep learning strategy (9)	.82	4.61	.91	.48***	.22***	.56***	.17***	.58***	.21***	.39***	.14**	1.00				
Surface learning strategy (10)	.84	4.98	.88	.26***	.34***	.21***	.32***	.24***	.35***	.29***	.32***	.16***	1.00			
Challenging tasks (11)	.82	4.94	.98	.37***	.30***	.45***	.18***	.44***	.21***	.34***	.19***	.43***	.29***	1.00		
Grade aspiration (12)	n/a	10.22	1.25	.14**	.15**	.19***	-.05	.20***	-.06	.19***	-.01	.21***	.00	.01	1.00	
Persistence (13)	.85	5.29	1.15	.48***	.36***	.49***	.13**	.53***	.12**	.36***	.09*	.39***	.43***	.40***	.25***	1.00

Note. n/a means not applicable (i.e., the scale only comprises one item).  
† *p* < .10. \* *p* < .05. \*\* *p* < .01. \*\*\* *p* < .001.



items to assess performance goals (e.g., "My goal is to perform better than the other students").

**Autonomous and controlled reasons for goal pursuit.** The same measure used in the prior study was used in this study, albeit "in my job" was replaced by "in my classes."

**Autonomous and controlled mastery and performance goal complexes.** Autonomous and controlled achievement goal complexes were operationalized in the same way as in the prior studies (i.e., by combining each goal statement with each reason statement): Four items (2 goal items  $\times$  2 reason items) assessed the autonomous mastery goal complex, eight items (2 goal items  $\times$  4 reason items) assessed the controlled mastery goal complex, four items (2 goal items  $\times$  2 reason items) assessed the autonomous performance goal complex, and eight items (2 goal items  $\times$  4 reason items) assessed the controlled performance goal complex.

**Deep and surface learning.** Kirby et al.'s (2003) Approaches to Learning at Work Questionnaire was adapted to the academic domain. Ten items assessed deep learning (e.g., "I spend a good deal of my spare time learning about things related to my classes") and 10 items assessed surface learning (e.g., "The best way for me to understand what technical terms mean is to remember the textbook definitions").

**Challenging tasks.** Preenen et al.'s (2011) six-item Challenging Assignments Scale was adapted to the academic domain to assess challenging tasks (e.g., "[In my classes I perform tasks] that are challenging").

**Grade aspiration.** McGregor and Elliot's (2002) single item measure was used to assess grade aspiration. Participants were asked to indicate "the minimum average grade that [they] would be satisfied with in [their] classes this semester" using a 12-point scale ranging from A to F (coded A = 12, A- = 11, B+ = 10 . . . , F = 1).

**Study persistence.** Elliot et al.'s (1999) four-item persistence subscale was used to assess study persistence (e.g., "When something that I am studying gets difficult, I spend extra time and effort trying to understand it").

## Results

**Overview.** We used the same analytical strategy used in Studies 2 and 3, albeit performance goals were included in the goal models. For each outcome variable, four models were built: the "goal-only" model (including mastery and performance goals; Model 1 in Tables 7 and 8), the "reason-only" model (including autonomous and controlled reasons; Model 2 in Tables 7 and 8), the "goal-and-reason" model (including mastery and performance goals *and* autonomous and controlled reasons; Model 3 in Tables 7 and 8), and the "goal complex" model (including achievement goals, reasons, *and* autonomous and controlled mastery and performance goal complexes; Model 4 in Tables 7 and 8).

**Preliminary analysis.** As in Studies 1–3, we conducted a preliminary analysis to examine potential covariates (sex, age, year at school) and order effects. None of the covariates attained significance ( $ps > .111$ ), excepting a negative association between sex and deep learning,  $B = -0.33 [-0.49, -0.17]$ ,  $p < .001$ , and between age and challenging tasks,  $B = -0.06 [-0.12, 0]$ ,  $p = .049$ . Although no order main effects were observed ( $ps > .116$ ), order interacted with performance goals in predicting persistence,  $B = -0.17 [-0.33, -0.01]$ ,  $p = .042$ . Again, as including these terms was neither theoretically relevant nor changed the pattern of results, they were not considered further.

## Main Analyses

**Deep learning and challenging tasks.** Table 7 presents the full set of results.

**"Goal-only" model.** In line with Hypothesis 1a, mastery goals were a positive predictor of deep learning,  $B = 0.35 [0.28, 0.42]$ ,  $p < .001$ , and challenging tasks,  $B = 0.25 [0.18, 0.33]$ ,  $p < .001$ .

**"Reason-only" model.** In line with Hypothesis 1b, autonomous reasons were a positive predictor of deep learning,  $B = 0.44 [0.38, 0.50]$ ,  $p < .001$ , and challenging tasks,  $B = 0.38 [0.30, 0.45]$ ,  $p < .001$ .

Table 7

*Study 4 (Deep Learning and Challenging Tasks): Coefficient Estimates and Effect Sizes for the Models Testing the Influence of Achievement Goals Alone (Model 1; "Goal-Only" Model), Autonomous and Controlled Reasons Alone (Model 2; "Reason-Only" Model), Achievement Goals and Reasons (Model 3; "Goal-and-Reason" Model), and Achievement Goals, Reasons, and Goal Complexes (Model 4; "Goal Complex" Model)*

	Deep learning strategies								Challenging tasks							
	Model 1		Model 2		Model 3		Model 4		Model 1		Model 2		Model 3		Model 4	
	<i>B</i>	$\eta_p^2$	<i>B</i>	$\eta_p^2$	<i>B</i>	$\eta_p^2$	<i>B</i>	$\eta_p^2$	<i>B</i>	$\eta_p^2$	<i>B</i>	$\eta_p^2$	<i>B</i>	$\eta_p^2$	<i>B</i>	$\eta_p^2$
Intercept	2.52	—	1.97	—	1.69	—	1.47	—	2.85	—	2.51	—	2.15	—	1.95	—
Mastery goals (MAP)	.35***	.19			.17***	.04	.08 <sup>†</sup>	.01	.25***	.09			.10*	.01	.05	—
Performance goals (PAP)	.04	—			-.02	—	-.09*	.01	.14***	.03			.09*	.01	.04	—
Autonomous reasons			.44***	.31	.34***	.14	.22***	.05			.38***	.19	.29***	.08	.21***	.03
Controlled reasons			.09**	.02	.09**	.02	.00	—			.12***	.02	.08*	.01	-.01	—
Autonomous MAP complex							.20***	.03							.15*	.01
Controlled MAP complex							.09	—							.04	—
Autonomous PAP complex							.13***	.03							.05	—
Controlled PAP complex							.00	—							.07	—

*Note.* Variables are not centered. ">" means that the predictive strength of mastery goals in Model 1 is significantly greater than the predictive strength of mastery goals in Model 3 (i.e., there is a significant reduction from Model 1 to Model 3). This is the case for the other model comparisons (i.e., Model 2 vs. 3, and Model 3 vs. 4) and variable (i.e., autonomous reasons) as well.

<sup>†</sup>  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

Table 8

Study 4 (Surface learning, Grade aspiration, and Study Persistence): Coefficient Estimates and Effect Sizes for the Models Testing the Influence of Achievement Goals Alone (Model 1; "Goal-Only" Model), Autonomous and Controlled Reasons Alone (Model 2; "Reason-Only" Model), Achievement Goals and Reasons (Model 3; "Goal-and-Reason" Model), and Achievement Goals, Reasons, and Goal Complexes (Model 4, "Goal Complex" Model)

	Surface learning strategies						Grade aspiration						Study persistence					
	Model 1		Model 2		Model 3		Model 4		Model 1		Model 2		Model 3		Model 4		Model 1	
	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$	B	$\eta^2_p$
Intercept	3.36	—	3.27	—	2.83	—	2.71	—	9.00	—	9.39	—	9.07	—	2.22	—	2.43	—
Mastery goals (MAP)	.12***	.02	.11*	.01	.12***	.01	.09†	.01	.11*	.01	.12***	.01	.15**	.02	.39***	.17	.23***	.04
Performance goals (PAP)	.19***	.07	.12***	.03	.12***	.01	.09*	.01	.12*	.01	.22***	.04	.18**	.02	.19***	.05	.16***	.04
Autonomous reasons			.14***	.04	.03	—	.02	—									.48***	.23
Controlled reasons			.23***	.10	.17***	.05	.05	—			-.07	—	-.13*	.01	.09*	.01	.29***	.07
Autonomous MAP complex							.05	—									.01	—
Controlled MAP complex							.05	—									.25***	.03
Autonomous PAP complex							.12	—									-.10	—
Controlled PAP complex							.02	—									.08†	.01
							.05	—									-.03	—

Note. Variables are not centered. " > " means that the predictive strength of mastery (performance) goals in Model 1 is significantly or marginally greater than the predictive strength of mastery (performance) goals in Model 3 (i.e., there is a significant or marginal reduction from Model 1 to Model 3); " = " means that the difference is not significant. This is the case for the other model comparisons (i.e., Model 2 vs. 3, and Model 3 vs. 4) and variables (i.e., performance goals and autonomous reasons) as well.

†  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

**"Goal-and-reason" model.** In line with Hypothesis 2a, mastery goals remained a positive predictor of deep learning,  $B = 0.17$  [0.09, 0.24],  $p < .001$ , and challenging tasks,  $B = 0.10$  [0.01, 0.18],  $p = .031$ . In line with Hypothesis 2b, autonomous reasons remained a positive predictor of deep learning,  $B = 0.34$  [0.26, 0.41],  $p < .001$ , and challenging tasks,  $B = 0.29$  [0.20, 0.37],  $p < .001$ .

In line with hypothesis 3a, the Monte Carlo method revealed that the reduction of the relations between mastery goals and both deep learning,  $B = 0.19$  [0.14, 0.24] (53% reduction), and challenging tasks,  $B = 0.16$  [0.11, 0.22] (63%), due to the inclusion of autonomous reasons were significant ( $Z_s \geq 5.85$ ,  $ps < .001$ ). In line with Hypothesis 3b, the reduction of the relations between autonomous reasons and both deep learning,  $B = 0.10$  [0.05, 0.14] (22%), and challenging tasks,  $B = 0.06$  [0.01, 0.11] (16%), due to the inclusion of mastery goals were significant ( $Z_s \geq 2.15$ ,  $ps \leq .032$ ).

**"Goal complex" model.** In line with Hypothesis 4, the autonomous mastery goal complex was a positive predictor of deep learning,  $B = 0.20$  [0.10, 0.31],  $p < .001$ , and challenging tasks,  $B = 0.15$  [0.02, 0.28],  $p = .023$ .

In line with Hypothesis 5a, the Monte Carlo method revealed that the reduction of the relations between mastery goals and both deep learning,  $B = 0.08$  [0.04, 0.13] (49%), and challenging tasks,  $B = 0.06$  [0.01, 0.11] (56%), due to the inclusion of the autonomous mastery goal complex were significant ( $Z_s \geq 2.24$ ,  $ps \leq .025$ ; mastery goals respectively became a marginal,  $p = .057$ , and a nonsignificant,  $p = .374$ , predictor). In line with Hypothesis 5b, the reduction of the relations between autonomous reasons and both deep learning,  $B = 0.08$  [0.04, 0.13] (27%), and challenging tasks,  $B = 0.06$  [0.01, 0.11] (22%), due to the inclusion of the autonomous mastery goal complex were significant ( $Z_s \geq 2.24$ ,  $ps \leq .025$ ; autonomous reasons remained a significant predictor in both instances,  $ps < .001$ ).

**Surface learning and grade aspiration.** Table 8 presents the full set of results.

**"Goal-only" model.** In line with Hypothesis 1a, performance goals were a positive predictor of surface learning,  $B = 0.19$  [0.13, 0.25],  $p < .001$ , and grade aspiration,  $B = 0.12$  [0.02, 0.21],  $p = .018$ .<sup>5</sup>

**"Goal-and-reason" model.** In line with Hypothesis 2a, performance goals remained a positive predictor of surface learning,  $B = 0.12$  [0.06, 0.19],  $p < .001$ , and grade aspiration,  $B = 0.15$  [0.05, 0.26],  $p = .004$ . Hypothesis 2b, 3a, and 3b were not formulated.

**"Goal complex" model.** Contrary to Hypothesis 4, the autonomous performance goal complex was not a positive predictor of surface learning,  $B = 0.02$  [-0.07, 0.10],  $p = .708$ ; in line with Hypothesis 4, the autonomous performance goal complex was a positive predictor of grade aspiration,  $B = 0.13$  [0, 0.27],  $p = .047$ .

Hypothesis 5a was not tested for surface learning, given the null result for the autonomous performance goal complex. In line with Hypothesis 5a, the Monte Carlo method revealed that the 36% reduction of the relation between performance goals and grade

<sup>5</sup> Thirty-eight participants did not provide an answer to the single-item grade aspiration scale; they were treated as missing values for this outcome variable.



aspiration due to the inclusion of the autonomous performance goal complex was significant,  $B = 0.05$ ,  $[0, 0.10]$  (although  $Z = 1.94$ ,  $p = .051$ ; performance goals became a nonsignificant predictor,  $p = .158$ ). Hypothesis 5b was not formulated.

**Persistence.** Table 8 presents the full set of results.

**“Goal-only” model.** In line with Hypothesis 1a, both mastery goals and performance goals were a positive predictor of study persistence,  $B = 0.39$   $[0.31, 0.47]$ ,  $p < .001$ , and  $B = 0.19$   $[0.11, 0.26]$ ,  $p < .001$ , respectively.

**“Reason-only” model.** In line with Hypothesis 1b, autonomous reasons were a positive predictor of study persistence,  $B = 0.48$   $[0.40, 0.57]$ ,  $p < .001$ .

**“Goal-and-reason” model.** In line with Hypothesis 2a, both mastery goals,  $B = 0.23$   $[0.13, 0.32]$ ,  $p < .001$ , and performance goals,  $B = 0.16$   $[0.08, 0.24]$ ,  $p < .001$ , remained a positive predictor of study persistence. In line with Hypothesis 2b, autonomous reasons remained a positive predictor of study persistence,  $B = 0.29$   $[0.19, 0.39]$ ,  $p < .001$ .

In line with hypothesis 3a, the Monte Carlo method revealed that the 42% reduction of the relation between mastery goals and study persistence due to the inclusion of autonomous reasons was significant,  $B = 0.16$   $[0.11, 0.22]$  ( $Z = 5.42$ ,  $p < .001$ ); the corresponding 11% reduction of the relation between performance goals and study persistence was marginal,  $B = 0.02$   $[0, 0.04]$  ( $Z = 1.77$ ,  $p = .077$ ). In line with Hypothesis 3b, the 31% reduction of the relation between autonomous reasons and study persistence due to the inclusion of mastery goals was significant,  $B = 0.13$   $[0.07, 0.19]$  ( $Z = 4.39$ ,  $p < .001$ ); the corresponding 6% reduction of the relation between autonomous reasons and study persistence due to the inclusion of performance goals was marginal,  $B = 0.02$   $[0, 0.04]$  ( $Z = 1.69$ ,  $p = .092$ ).

**“Goal complex” model.** In line with Hypothesis 4, the autonomous mastery goal complex was a positive predictor of study persistence,  $B = 0.25$   $[0.11, 0.40]$ ,  $p < .001$ , and the autonomous performance goal complex was a marginally significant positive predictor,  $B = 0.08$   $[-0.01, 0.18]$ ,  $p = .092$ .

In line with Hypothesis 5a, the Monte Carlo method revealed that the 45% reduction of the relation between mastery goals and study persistence due to the inclusion of the autonomous mastery goal complex was significant,  $B = 0.10$   $[0.04, 0.16]$  ( $Z = 3.36$ ,  $p < .001$ ; mastery goals remained a positive predictor,  $p = .035$ ). The 18% reduction of the relation between performance goals and study persistence due to the inclusion of the autonomous performance goal complex was marginal,  $B = 0.03$   $[0, 0.07]$  ( $Z = 1.66$ ,  $p = .098$ ). In line with Hypothesis 5b, the 39% reduction of the relation between autonomous reasons and study persistence due to the inclusion of the autonomous mastery goal complex was significant,  $B = 0.10$   $[0.04, 0.16]$  ( $Z = 3.36$ ,  $p < .001$ ; autonomous reasons remained a positive predictor,  $p = .009$ ); the corresponding 4% reduction due to the inclusion of the autonomous performance goal complex was nonsignificant,  $B = 0.10$   $[0, 0.23]$  ( $Z = 1.13$ ,  $p = .260$ ).

## Discussion

Replicating Study 3’s findings, mastery goals and autonomous reasons accounted for variance in deep learning and challenging tasks when tested separately or simultaneously (with the predictive strength of each being diminished). Moreover, the autonomous

mastery goal complex explained additional variance in deep learning and challenging tasks, and diminished the predictive strength of both mastery goals and autonomous reasons.

Extending Study 3’s findings, performance goals accounted for variance in surface learning and grade aspiration, when testing goals and reasons separately or simultaneously. Moreover, the autonomous performance goal complex explained additional variance in grade aspiration, and diminished the predictive strength of performance goals. The autonomous performance goal complex did not explain additional variance in surface learning.

Further extending Study 3’s findings, mastery goals, performance goals, and autonomous reasons accounted for variance in study persistence when testing goals and reasons separately or simultaneously (with the predictive strength of each being diminished). Moreover, the autonomous mastery and performance goal complexes explained additional variance in persistence, and diminished the predictive strength of mastery goals, performance goals, and autonomous reasons. The reductions of the influence of performance goals and the influence of the autonomous performance goal complex only attained marginal significance.

## General Discussion

Although research on achievement goals and reasons has only recently commenced, there has been a growing interest in studying the SDT-derived reasons connected to achievement goals (see Vansteenkiste, Lens, et al., 2014). The findings from this work have often been interpreted as indicating that the influence of achievement goals on beneficial outcomes is reducible to the influence of reasons. In the present research, we developed a systematic approach to studying goals, reasons, and goal complexes, and utilized this approach to clearly differentiate between the influence of achievement goals, autonomous and controlled reasons, and achievement goal complexes. Our results revealed that all three types of variables accounted for independent variance in experiential and self-regulated learning outcomes.

## Summary of Findings

First, we documented the *separate* influence of mastery goals and autonomous reasons for goal pursuit. On the one hand, mastery goals were found to be a positive predictor of beneficial experiential (satisfaction, interest, and positive emotion) and self-regulated learning (deep learning, interpersonal help-seeking, challenging tasks, and persistence) outcomes. This replicates basic findings from the achievement goal literature, showing that mastery goals enhance the subjective value of the achievement activity and foster interest-based learning processes (Daniels et al., 2009). On the other hand, autonomous reasons were found to be a positive predictor of the same beneficial outcomes. This replicates basic findings from the SDT literature, showing that reasons involving the self-endorsement of one’s actions enhance task enjoyment and facilitate growth (Deci et al., 1991).

Second, we documented the *simultaneous* influence of mastery goals and autonomous reasons for goal pursuit. On the one hand, both mastery goals and autonomous reasons were found to explain independent variance in most of the beneficial experiential (interest and positive emotion) and self-regulated learning (deep learning, challenging tasks, and persistence) outcomes. This illustrates



that mastery goals and autonomous reasons are *distinct* motivational constructs, presumably having similar influences via different processes (Dysvik & Kuvaas, 2010). On the other hand, the predictive strength of mastery goals and autonomous reasons for these same outcomes were each found to be diminished when controlling for the other. This illustrates that mastery goals and autonomous reasons are *overlapping* motivational constructs, both pertaining to an internal investment in the value of learning (Elliot, & Church, 1997). However, controlling for mastery goals eliminated the link between autonomous reasons and interpersonal help-seeking, whereas controlling for autonomous reasons eliminated the link between mastery goals and satisfaction. This suggests that the influence of reasons may outweigh the influence of goals for some outcomes, but that the influence of goals may outweigh the influence of reasons for other outcomes.

Third, we documented the influence of the autonomous mastery goal complex *together* with mastery goals and autonomous reasons for goal pursuit. On the one hand, the autonomous mastery goal complex was found to explain incremental variance in all of the beneficial experiential outcomes (interest, satisfaction, and positive emotion) and most of the beneficial self-regulated learning outcomes (i.e., deep learning, challenging tasks, and persistence). This indicates that the autonomous mastery goal complex is more than the mere sum of a mastery goal and autonomous reasons: Autonomous reasons may give deeper psychological meaning to the mastery goal, and the mastery goal may then foster a pleasurable, interest-driven approach to learning (Ryan & Deci, 2006). On the other hand, the predictive strength of mastery goals and autonomous reasons regarding these same outcomes were each found to be diminished when controlling for the autonomous mastery goal complex. This is likely due to measurement redundancy: Mastery goals and autonomous reasons were each measured (at least) two times, first as a “pure” goal or a “pure” reason, and second as a part of the autonomous mastery goal complex. However, for many outcomes, mastery goals and autonomous reasons still explained residual variance after controlling for the autonomous mastery goal complex. Hence, it appears that mastery goals in and of themselves (or, perhaps more accurately, mastery goals *energized* by reasons not captured by the goal complexes examined herein) and autonomous reasons in and of themselves (or, perhaps more accurately, autonomous reasons *directed* by aims not captured by the goal complexes examined herein) each have remaining, substantive predictive utility.

Fourth, we also documented the influence of performance goals and performance goal complexes. Performance goals were found to be a positive predictor of surface learning, grade aspiration, and study persistence, even after controlling for reasons for goal pursuit. Moreover, the autonomous performance goal complex explained incremental variance in grade aspiration and study persistence, resulting in the diminution of the predictive strength of both performance goals (for grade aspiration) and autonomous reasons (for persistence). In the same way as for mastery goals, these results show that performance goal content matters, and does so in two ways: The influence of performance goals is not reducible to the influence of reasons, and the pattern of results associated with the autonomous performance goal complex differs from that associated with the autonomous mastery goal complex.

Fifth, in ancillary analyses we observed the influence of controlled achievement goal complexes. In nearly all instances, con-

trolled achievement goal complexes did not explain incremental variance in the beneficial experiential and self-regulated learning outcomes (the lone exception—of 22 instances—being controlled mastery goal complexes and deep learning in Study 2). Mastery and performance goals do *not* seem to provide supplementary benefits when combined with controlled reasons, which is consistent with research showing that endorsing these goals for self-presentation purposes (a form of controlled reason) lessens or eliminates their positive influence (Dompnier, Darnon, & Butera, 2013; Smeding et al., 2015).

### Both Goals and Reasons Are Needed for a Full Account of Motivation

The present research echoes a past controversy in the motivation literature. SDT researchers have long distinguished between *intrinsic* (e.g., growth, relationships, community) and *extrinsic* (e.g., wealth, fame, image) goal content (for a review, see Vansteenkiste, Lens, & Deci, 2006). Intrinsic goals tend to predict beneficial outcomes, whereas extrinsic goals tend to predict detrimental outcomes (Kasser & Ryan, 1996). In the late 1990s, the relation between intrinsic goals and a self-regulation outcome (self-actualization) was found to be eliminated when partialing out the influence of the autonomous and controlled reasons connected to these goals (Carver & Baird, 1998). The authors interpreted this finding as suggesting that “it often matters more why a goal is being pursued than what the goal is” (p. 292). Later, the relation between extrinsic goals and an experiential outcome (well-being) was also found to be eliminated when controlling for the autonomous-like (i.e., freedom of action motives) and controlled-like (i.e., appearing worthy in others’ eyes) reasons connected to these goals (Srivastava, Locke, & Bartol, 2001). Here too the conclusion was reached that the predictive utility of goals is negligible once reasons are considered.

However, Sheldon, Ryan, Deci, and Kasser (2004) critiqued the aforementioned research, highlighting that goal assessment was confounded with reason assessment. After refining the methodology of the prior work, Sheldon et al. (2004) demonstrated that both goal content (i.e., intrinsic vs. extrinsic goals) and goal motives (i.e., autonomous vs. controlled reasons) made significant and independent contributions to psychological well-being. They came to the conclusion that neither the directive focus of goals nor the dynamic processes underlying goals was more critical than the other (for similar work showing that both goal content and reasons are important to understand outcomes in the exercise domain, see Sebire, Standage, & Vansteenkiste, 2009).

Similar reasoning applies to the emerging research on goal complexes within the achievement domain. In prior work, the relation between achievement goals and a series of achievement-relevant outcomes (e.g., positive emotion, engagement, persistence) was found to be eliminated when partialing out the influence of the autonomous reasons connected to these goals (see Gillet et al., 2015; Vansteenkiste, Mouratidis, et al., 2010; Vansteenkiste, Smeets, et al., 2010). Because this prior work did not include “pure reason” assessments, we believe that this type of reduction should be interpreted with caution. Indeed, our findings indicate that the influence of *achievement goal content* is not reducible to the influence of *achievement goal motives*. The influence of achievement goals is not unilaterally exceeded by the



influence of reasons, and the influence of achievement goal complexes both depends on the type of goal and the type of reason they encompass. As such, it is best for scholars to resist “either-or” perspectives on achievement motivation: Not only do reasons for goal pursuit matter, but the goals themselves matter as well. Thus, we concur with Vansteenkiste, Mouratidis, et al.’s (2014) statement that “reasons [should] not [be] meant to replace the achievement goals themselves” (p. 142).

### Short-Term and Long-Term Research Directions

We believe that a clear conceptual and empirical disentanglement of achievement goals and reasons brings a fresh, exciting, and generative perspective to the achievement goal literature. In the short term, researchers may consider adopting a cumulative approach that involves further investigating the influence of achievement goals, reasons, and achievement goal complexes on achievement-relevant outcomes. Specifically, researchers may focus on other achievement goals (e.g., avoidance-based goals; see Gillet et al., 2015), non SDT-derived reasons (e.g., achievement motives, Elliot, 1999; social motivation, Ryan & Shim, 2008; competitive motives, Murayama & Elliot, 2012), unusual goal complexes (e.g., formed upon the adoption of maladaptive goals and adaptive reasons, such as the autonomous performance-avoidance complex; see Heidemeier & Wiese, 2014), and/or a wider range of outcomes (e.g., beneficial *and* detrimental; see Senko, 2016).

In the long-term, researchers may consider adopting a more comprehensive approach that involves moving beyond comparison of the influence of achievement goals, reasons, and achievement goal complexes. Conceptualizing and operationalizing achievement goal complexes raise two important, intertwined issues that need to be addressed in future work: Complexity and ecological validity. Regarding complexity, the most elaborate achievement goal framework encompasses  $3 \times 2$  achievement goals (i.e., task-, self-, and other-based standards crossed with approach and avoidance; Elliot, Murayama, & Pekrun, 2011), and the self-determination framework encompasses five main types of reasons (i.e., extrinsic reasons with external, introjected, identified, or integrated regulation, and intrinsic reasons; Ryan & Deci, 2000). Fully integrating these frameworks would result in  $3 \times 2 \times 5 = 30$  possible achievement goal complexes, which are clearly too many to rigorously study at the same time. As such, it is important for researchers to select a subset of achievement goals and reasons in any given investigation to avoid overtaxing participants with a large number of related and (seemingly) redundant questions (which would undoubtedly yield poor quality data).

Regarding ecological validity, researchers may consider which achievement goal complexes are more commonly encountered in real-life achievement settings. It is known that mastery-approach, performance-approach, and performance-avoidance are spontaneously generated by participants (in their own words) in open-ended questions or semistructured interviews (Lee & Bong, 2016; Levy, Kaplan, & Patrick, 2004; Urdan, 2004b). However, little is known about the spontaneously generated reasons behind mastery-approach, performance-approach, and performance-avoidance goals (for an exception, see Urdan & Mestas, 2006). Future research would benefit from using *inductive* methods to determine the most prevalent achievement goal-reason combinations

(and whether SDT or some other approach or approaches to motivation is/are best suited to conceptualize these achievement goal complexes) and using *deductive* methods to estimate their consequences for achievement-relevant outcomes. Such a mixed method research program (see Johnson & Onwuegbuzie, 2004) would help motivation scientists to focus their conceptual attention and empirical effort on variables of foremost practical significance.

### Limitations

The limitations of our work should be acknowledged. First, the present studies were correlational and relied on single-session data collections. Hence, we cannot establish the causal nature of the motivation-to-outcome relations. Subsequent research using prospective methods is needed to acquire more precise insight into these dynamics. For instance, motivational and outcome variables could be assessed at different times (as in Harackiewicz et al., 1997) or a longitudinal design could be employed (as in Daniels et al., 2009).

Second, mastery goals and autonomous reasons were moderately to highly correlated ( $r \approx .60$ ), as in past research (e.g., Katz et al., 2008). That is, the two motivational constructs are multicollinear, suggesting that mastery goals are primarily pursued for autonomous reasons (see Senko & Tropicano, 2016). However, it should be noted that multicollinearity is not a violation of the assumptions of ordinary least squares estimation (Freund & Littell, 2000). Multiple regression analysis has enabled us to estimate the *unique* variance explained by mastery goals, after removing the *shared* variance associated with autonomous reasons (and vice versa). The only risk with multicollinearity stems from a lack of information in the data (e.g., participants with high mastery goals and low autonomous reasons are unusual; see Brambor, Clark, & Golder, 2006). In this regard, multicollinearity may have increased the probability of Type II error (false negative) but not that of Type I error (false positive; see Mason & Perreault Jr, 1991).

Third, the assessment of our main theoretical constructs, namely mastery goals, autonomous reasons, and beneficial outcomes, may be subject to social desirability (see Darnon, Dompnier, Delmas, Pulfrey, & Butera, 2009; Lepper, Corpus, & Iyengar, 2005). Thus, the link between these constructs might be partially explained by covarying interindividual differences in self-presentation. However, it is important to note that such impression-management issues cannot account for the robust finding that both achievement goals and reasons have independent predictive utility. Nevertheless, subsequent research would benefit from controlling for social desirability and incorporating behavioral measures falling outside the categories of the variables studied in the present article (e.g., achievement, see Senko, Hulleman, & Harackiewicz, 2011).

Fourth, our studies were based on U.S. samples. The levels of both achievement goals and self-determined motivation have been found to vary somewhat across culture (Chirkov & Ryan, 2001; Dekker & Fischer, 2008), as have predictive patterns for achievement goals (Zan, Xiang, Louis, Jianmin, & YunPeng, 2008; see Chirkov, 2009 on autonomous motivation, which may have more universal predictive power). Given these cross-cultural differences, research is needed to test the predictive utility of achievement goals, reasons, and achievement goal complexes in a broader array of countries.



## Conclusion

The achievement goals approach to achievement motivation identifies a number of possible *goal contents* in competence-relevant contexts that vary according to how competence is defined and valenced (Elliot et al., 2011), whereas SDT designates a continuum of possible *goal motives* ranging from autonomous to controlled (Deci & Ryan, 2000). Our research herein suggests that these two frameworks should be thought of in integrative rather than comparative terms: Achievement goals, reasons for goal pursuit, and achievement goal complexes all make independent contributions to experiential and self-regulated learning outcomes in achievement settings. In our view, conceptualizing, operationalizing, and empirically analyzing both the direction and energization of goal striving using both of these theoretical frameworks offers the most promising avenue for a full and complete account of competence motivation.

## References

- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261–271. <http://dx.doi.org/10.1037/0022-0663.84.3.261>
- Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology, 80*, 260–267. <http://dx.doi.org/10.1037/0022-0663.80.3.260>
- Baranik, L. E., Stanley, L. J., Bynum, B. H., & Lance, C. E. (2010). Examining the construct validity of mastery-avoidance achievement goals: A meta-analysis. *Human Performance, 23*, 265–282. <http://dx.doi.org/10.1080/08959285.2010.488463>
- Benita, M., Roth, G., & Deci, E. L. (2014). When are mastery goals more adaptive? It depends on experiences of autonomy support and autonomy. *Journal of Educational Psychology, 106*, 258–267. <http://dx.doi.org/10.1037/a0034007>
- Brambor, T., Clark, W. R., & Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis, 14*, 63–82. <http://dx.doi.org/10.1093/pan/mpi014>
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5. <http://dx.doi.org/10.1177/1745691610393980>
- Carver, C. S., & Baird, E. (1998). The American dream revisited: Is it what you want or why you want it that matters? *Psychological Science, 9*, 289–292. <http://dx.doi.org/10.1111/1467-9280.00057>
- Chirkov, V. I. (2009). A cross-cultural analysis of autonomy in education. A self-determination theory perspective. *Theory and Research in Education, 7*, 253–262. <http://dx.doi.org/10.1177/1477878509104330>
- Chirkov, V. I., & Ryan, R. M. (2001). Parent and teacher autonomy-support in Russian and U.S. adolescents' common effects on well-being and academic motivation. *Journal of Cross-Cultural Psychology, 32*, 618–635. <http://dx.doi.org/10.1177/0022022101032005006>
- Daniels, L. M., Stupnisky, R. H., Pekrun, R., Haynes, T. L., Perry, R. P., & Newall, N. E. (2009). A longitudinal analysis of achievement goals: From affective antecedents to emotional effects and achievement outcomes. *Journal of Educational Psychology, 101*, 948–963. <http://dx.doi.org/10.1037/a0016096>
- Darnon, C., Dompnier, B., Delmas, F., Pulfrey, C., & Butera, F. (2009). Achievement goal promotion at university: Social desirability and social utility of mastery and performance goals. *Journal of Personality and Social Psychology, 96*, 119–134. <http://dx.doi.org/10.1037/a0012824>
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum Press. <http://dx.doi.org/10.1007/978-1-4899-2271-7>
- Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*, 227–268. [http://dx.doi.org/10.1207/S15327965PLI1104\\_01](http://dx.doi.org/10.1207/S15327965PLI1104_01)
- Deci, E. L., & Ryan, R. M. (2008). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology, 49*, 182–185. <http://dx.doi.org/10.1037/a0012801>
- Deci, E. L., & Ryan, R. M. (2016). Optimizing students' motivation in the era of testing and pressure: A self-determination theory perspective. In C. W. Liu, K. J. C. Wang, & M. R. Ryan (Eds.), *Building autonomous Learners: Perspectives from research and practice using self-determination theory* (pp. 9–29). Singapore: Springer. [http://dx.doi.org/10.1007/978-981-287-630-0\\_2](http://dx.doi.org/10.1007/978-981-287-630-0_2)
- Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist, 26*, 325–346. <http://dx.doi.org/10.1080/00461520.1991.9653137>
- Dekker, S., & Fischer, R. (2008). Cultural differences in academic motivation goals: A meta-analysis across 13 societies. *The Journal of Educational Research, 102*, 99–110. <http://dx.doi.org/10.3200/JOER.102.2.99-110>
- Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71–75. [http://dx.doi.org/10.1207/s15327752jpa4901\\_13](http://dx.doi.org/10.1207/s15327752jpa4901_13)
- Diseth, Å. (2011). Self-efficacy, goal orientations and learning strategies as mediators between preceding and subsequent academic achievement. *Learning and Individual Differences, 21*, 191–195. <http://dx.doi.org/10.1016/j.lindif.2011.01.003>
- Diseth, Å., & Samdal, O. (2014). Autonomy support and achievement goals as predictors of perceived school performance and life satisfaction in the transition between lower and upper secondary school. *Social Psychology of Education, 17*, 269–291. <http://dx.doi.org/10.1007/s11218-013-9244-4>
- Dompnier, B., Darnon, C., & Butera, F. (2009). Faking the desire to learn: A clarification of the link between mastery goals and academic achievement. *Psychological Science, 20*, 939–943. <http://dx.doi.org/10.1111/j.1467-9280.2009.02384.x>
- Dompnier, B., Darnon, C., & Butera, F. (2013). When performance-approach goals predict academic achievement and when they do not: A social value approach. *British Journal of Social Psychology, 52*, 587–596. <http://dx.doi.org/10.1111/bjso.12025>
- Donche, V., De Maeyer, S., Coertjens, L., Van Daal, T., & Van Petegem, P. (2013). Differential use of learning strategies in first-year higher education: The impact of personality, academic motivation, and teaching strategies. *The British Journal of Educational Psychology, 83*, 238–251. <http://dx.doi.org/10.1111/bjep.12016>
- Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040–1048. <http://dx.doi.org/10.1037/0003-066X.41.10.1040>
- Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia, PA: Psychology Press.
- Dysvik, A., & Kuvaas, B. (2010). Exploring the relative and combined influence of mastery-approach goals and work intrinsic motivation on employee turnover intention. *Personnel Review, 39*, 622–638. <http://dx.doi.org/10.1108/00483481011064172>
- Dysvik, A., & Kuvaas, B. (2013). Intrinsic and extrinsic motivation as predictors of work effort: The moderating role of achievement goals. *British Journal of Social Psychology, 52*, 412–430. <http://dx.doi.org/10.1111/j.2044-8309.2011.02090.x>
- Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34*, 169–189. [http://dx.doi.org/10.1207/s15326985ep3403\\_3](http://dx.doi.org/10.1207/s15326985ep3403_3)



- Elliot, A. J. (2005). A conceptual history of the achievement goal construct. In A. J. Elliot & C. Dweck (Eds.), *Handbook of competence and motivation* (pp. 52–72). New York, NY: Guilford Press.
- Elliot, A. J. (2006). The hierarchical model of approach-avoidance motivation. *Motivation and Emotion*, 30, 111–116. <http://dx.doi.org/10.1007/s11031-006-9028-7>
- Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology*, 72, 218–232. <http://dx.doi.org/10.1037/0022-3514.72.1.218>
- Elliot, A. J., & Fryer, J. W. (2008). The goal concept in psychology. In J. Shah & W. Gardner (Eds.), *Handbook of motivational science* (pp. 235–250). New York, NY: Guilford Press.
- Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology*, 80, 501–519. <http://dx.doi.org/10.1037/0022-3514.80.3.501>
- Elliot, A. J., McGregor, H. A., & Gable, S. (1999). Achievement goals, study strategies, and exam performance: A mediational analysis. *Journal of Educational Psychology*, 91, 549–563. <http://dx.doi.org/10.1037/0022-0663.91.3.549>
- Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology*, 100, 613–628. <http://dx.doi.org/10.1037/0022-0663.100.3.613>
- Elliot, A. J., Murayama, K., & Pekrun, R. (2011). A 3 × 2 achievement goal model. *Journal of Educational Psychology*, 103, 632–648. <http://dx.doi.org/10.1037/a0023952>
- Elliot, A. J., & Thrash, T. M. (2001). Achievement goals and the hierarchical model of achievement motivation. *Educational Psychology Review*, 13, 139–156. <http://dx.doi.org/10.1023/A:1009057102306>
- Fisher, C. D., Minbashian, A., Beckmann, N., & Wood, R. E. (2013). Task appraisals, emotions, and performance goal orientation. *Journal of Applied Psychology*, 98, 364–373. <http://dx.doi.org/10.1037/a0031260>
- Freud, R. J., & Littell, R. C. (2000). *SAS system for regression*. Cary, NC: SAS Institute.
- Gagné, M., & Deci, E. L. (2005). Self-determination theory and work motivation. *Journal of Organizational Behavior*, 26, 331–362. <http://dx.doi.org/10.1002/job.322>
- Gagné, M., Forest, J., Gilbert, M. H., Aubé, C., Morin, E., & Malorni, A. (2010). The Motivation at Work Scale: Validation evidence in two languages. *Educational and Psychological Measurement*, 70, 628–646. <http://dx.doi.org/10.1177/0013164409355698>
- Gaudreau, P. (2012). Goal self-concordance moderates the relationship between achievement goals and indicators of academic adjustment. *Learning and Individual Differences*, 22, 827–832. <http://dx.doi.org/10.1016/j.lindif.2012.06.006>
- Gaudreau, P., & Braaten, A. (2016). Achievement goals and their underlying goal motivation: Does it matter why sport participants pursue their goals? *Psychologica Belgica*, 56, 244–268. <http://dx.doi.org/10.5334/pb.266>
- Gillet, N., Lafrenière, M. A. K., Huyghebaert, T., & Fouquereau, E. (2015). Autonomous and controlled reasons underlying achievement goals: Implications for the 3 × 2 achievement goal model in educational and work settings. *Motivation and Emotion*, 39, 858–875. <http://dx.doi.org/10.1007/s11031-015-9505-y>
- Gillet, N., Lafrenière, M. A. K., Vallerand, R. J., Huart, I., & Fouquereau, E. (2014). The effects of autonomous and controlled regulation of performance-approach goals on well-being: A process model. *British Journal of Social Psychology*, 53, 154–174. <http://dx.doi.org/10.1111/bjso.12018>
- Harackiewicz, J. M., Barron, K. E., Carter, S. M., Lehto, A. T., & Elliot, A. J. (1997). Predictors and consequences of achievement goals in the college classroom: Maintaining interest and making the grade. *Journal of Personality and Social Psychology*, 73, 1284–1295. <http://dx.doi.org/10.1037/0022-3514.73.6.1284>
- Heidemeier, H., & Wiese, B. S. (2014). Achievement goals and autonomy: How person—Context interactions predict effective functioning and well-being during a career transition. *Journal of Occupational Health Psychology*, 19, 18–31. <http://dx.doi.org/10.1037/a0034929>
- Holman, D., Epitropaki, O., & Fernie, S. (2001). Understanding learning strategies in the workplace: A factor analytic investigation. *Journal of Occupational and Organizational Psychology*, 74, 675–681. <http://dx.doi.org/10.1348/096317901167587>
- Horton, J. J., & Chilton, L. B. (2010). The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce* (pp. 209–218). New York, NY: ACM. Retrieved from <https://arxiv.org/pdf/1001.0627.pdf>
- Huang, C. (2011). Achievement goals and achievement emotions: A meta-analysis. *Educational Psychology Review*, 23, 359–388. <http://dx.doi.org/10.1007/s10648-011-9155-x>
- Huang, C. (2016). Achievement goals and self-efficacy: A meta-analysis. *Educational Research Review*, 19, 119–137. <http://dx.doi.org/10.1016/j.edurev.2016.07.002>
- Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin*, 136, 422–449. <http://dx.doi.org/10.1037/a0018947>
- Janssen, O., & Van Yperen, N. W. (2004). Employees' goal orientations, the quality of leader-member exchange, and the outcomes of job performance and job satisfaction. *Academy of Management Journal*, 47, 368–384. <http://dx.doi.org/10.2307/20159587>
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33, 14–26. <http://dx.doi.org/10.3102/0013189X033007014>
- Kaplan, A., & Maehr, M. L. (2007). The contributions and prospects of goal orientation theory. *Educational Psychology Review*, 19, 141–184. <http://dx.doi.org/10.1007/s10648-006-9012-5>
- Karabenick, S. A. (2004). Perceived achievement goal structure and college student help seeking. *Journal of Educational Psychology*, 96, 569–581. <http://dx.doi.org/10.1037/0022-0663.96.3.569>
- Kasser, T., & Ryan, R. M. (1996). Further examining the American dream: Differential correlates of intrinsic and extrinsic goals. *Personality and Social Psychology Bulletin*, 22, 80–87. <http://dx.doi.org/10.1177/0146167296223006>
- Katz, I., Assor, A., & Kanat-Maymon, Y. (2008). A projective assessment of autonomous motivation in children: Correlational and experimental evidence. *Motivation and Emotion*, 32, 109–119. <http://dx.doi.org/10.1007/s11031-008-9086-0>
- Kirby, J. R., Knapper, C. K., Evans, C. J., Carty, A. E., & Gadula, C. (2003). Approaches to learning at work and workplace climate. *International Journal of Training and Development*, 7, 31–52. <http://dx.doi.org/10.1111/1468-2419.00169>
- Kusurkar, R. A., Ten Cate, T. J., Vos, C. M. P., Westers, P., & Croiset, G. (2013). How motivation affects academic performance: A structural equation modelling analysis. *Advances in Health Sciences Education*, 18, 57–69. <http://dx.doi.org/10.1007/s10459-012-9354-3>
- Lam, C. F., & Gurland, S. T. (2008). Self-determined work motivation predicts job outcomes, but what predicts self-determined work motivation? *Journal of Research in Personality*, 42, 1109–1115. <http://dx.doi.org/10.1016/j.jrp.2008.02.002>
- Lee, M., & Bong, M. (2016). In their own words: Reasons underlying the achievement striving of students in schools. *Journal of Educational Psychology*, 108, 274–294. <http://dx.doi.org/10.1037/edu0000048>
- Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and



- academic correlates. *Journal of Educational Psychology*, 97, 184–196. <http://dx.doi.org/10.1037/0022-0663.97.2.184>
- Levy, I., Kaplan, A., & Patrick, H. (2004). Early adolescents' achievement goals, social status, and attitudes towards cooperation with peers. *Social Psychology of Education*, 7, 127–159. <http://dx.doi.org/10.1023/B:SPOE.0000018547.08294.b6>
- Lewin, K. (1951). *Field theory in social science: Selected theoretical papers*. New York, NY: Harper & Row.
- Mackinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39, 99–128. [http://dx.doi.org/10.1207/s15327906mbr3901\\_4](http://dx.doi.org/10.1207/s15327906mbr3901_4)
- Maehr, M. L., & Nicholls, J. G. (1980). Culture and achievement motivation: A second look. In N. Warren (Ed.), *Studies in cross-cultural psychology* (Vol. 3, pp. 221–267). New York, NY: Academic Press.
- Mason, C. H., & Perreault, W. D., Jr. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*, 28, 268–280. <http://dx.doi.org/10.2307/3172863>
- McClelland, D. C. (1985). *Human motivation*. Glenview, IL: Scott Foresman.
- McGregor, H. A., & Elliot, A. J. (2002). Achievement goals as predictors of achievement-relevant processes prior to task engagement. *Journal of Educational Psychology*, 94, 381–395. <http://dx.doi.org/10.1037/0022-0663.94.2.381>
- Meece, J. L., Anderman, E. M., & Anderman, L. H. (2006). Classroom goal structure, student motivation, and academic achievement. *Annual Review of Psychology*, 57, 487–503. <http://dx.doi.org/10.1146/annurev.psych.56.091103.070258>
- Michou, A., Vansteenkiste, M., Mouratidis, A., & Lens, W. (2014). Enriching the hierarchical model of achievement motivation: Autonomous and controlling reasons underlying achievement goals. *The British Journal of Educational Psychology*, 84, 650–666. <http://dx.doi.org/10.1111/bjep.12055>
- Murayama, K., & Elliot, A. J. (2012). The competition-performance relation: A meta-analytic review and test of the opposing processes model of competition and performance. *Psychological Bulletin*, 138, 1035–1070. <http://dx.doi.org/10.1037/a0028324>
- Murray, H. (1938). *Explorations in personality*. New York, NY: Oxford University Press.
- Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review*, 91, 328–346. <http://dx.doi.org/10.1037/0033-295X.91.3.328>
- Nicholls, J. G. (1989). *The competitive ethos and democratic education*. Cambridge, MA: Harvard University Press.
- Ntoumanis, N., Healy, L. C., Sedikides, C., Duda, J., Stewart, B., Smith, A., & Bond, J. (2014). When the going gets tough: The “why” of goal striving matters. *Journal of Personality*, 82, 225–236. <http://dx.doi.org/10.1111/jopy.12047>
- Ozdemir Oz, A., Lane, J. F., & Michou, A. (2015). Autonomous and controlling reasons underlying achievement goals during task engagement: Their relation to intrinsic motivation and cheating. *Educational Psychology*. Advance online publication.
- Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18, 315–341. <http://dx.doi.org/10.1007/s10648-006-9029-9>
- Pekrun, R., Elliot, A. J., & Maier, M. A. (2006). Achievement goals and discrete achievement emotions: A theoretical model and prospective test. *Journal of Educational Psychology*, 98, 583–597. <http://dx.doi.org/10.1037/0022-0663.98.3.583>
- Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research*, 31, 459–470. [http://dx.doi.org/10.1016/S0883-0355\(99\)00015-4](http://dx.doi.org/10.1016/S0883-0355(99)00015-4)
- Pintrich, P. R., & Garcia, T. (1991). Student goal orientation and self-regulation in the college classroom. In M. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement: Goals and self-regulatory processes* (Vol. 7, pp. 371–402). Greenwich, CT: JAI.
- Preenen, P. T. Y., de Pater, I. E., van Vianen, A. E. M., & Keijzer, L. (2011). Managing voluntary turnover through challenging assignments. *European Journal of Work and Organizational Psychology*, 23, 48–61. <http://dx.doi.org/10.1080/1359432X.2012.702420>
- Ratelle, C. F., Guay, F., Vallerand, R. J., Larose, S., & Senécal, C. (2007). Autonomous, controlled, and amotivated types of academic motivation: A person-oriented analysis. *Journal of Educational Psychology*, 99, 734–746. <http://dx.doi.org/10.1037/0022-0663.99.4.734>
- Retelsdorf, J., Butler, R., Streblow, L., & Schiefele, U. (2010). Teachers' goal orientations for teaching: Associations with instructional practices, interest in teaching, and burnout. *Learning and Instruction*, 20, 30–46. <http://dx.doi.org/10.1016/j.learninstruc.2009.01.001>
- Ryan, A. M., & Shim, S. S. (2008). An exploration of young adolescents' social achievement goals and social adjustment in middle school. *Journal of Educational Psychology*, 100, 672–687. <http://dx.doi.org/10.1037/0022-0663.100.3.672>
- Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, 42, 450–461. <http://dx.doi.org/10.1037/0022-3514.43.3.450>
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78. <http://dx.doi.org/10.1037/0003-066X.55.1.68>
- Ryan, R. M., & Deci, E. L. (2006). Self-regulation and the problem of human autonomy: Does psychology need choice, self-determination, and will? *Journal of Personality*, 74, 1557–1585. <http://dx.doi.org/10.1111/j.1467-6494.2006.00420.x>
- Ryan, R. M., & Powelson, C. L. (1991). Autonomy and relatedness as fundamental to motivation and education. *Journal of Experimental Education*, 60, 49–66. <http://dx.doi.org/10.1080/00220973.1991.10806579>
- Sebire, S. J., Standage, M., & Vansteenkiste, M. (2009). Examining intrinsic versus extrinsic exercise goals: Cognitive, affective, and behavioral outcomes. *Journal of Sport & Exercise Psychology*, 31, 189–210. <http://dx.doi.org/10.1123/jsep.31.2.189>
- Senko, C. (2016). Achievement goal theory: A story of early promises, eventual discords, and future possibilities. In K. Wentzel & D. Miele (Eds.), *Handbook of motivation at school* (Vol. 2, pp. 75–95). New York, NY: Routledge.
- Senko, C., Hama, H., & Belmonte, K. (2013). Achievement goals, study strategies, and achievement: A test of the “learning agenda” framework. *Learning and Individual Differences*, 24, 1–10. <http://dx.doi.org/10.1016/j.lindif.2012.11.003>
- Senko, C., Hulleman, C. S., & Harackiewicz, J. M. (2011). Achievement goal theory at the crossroads: Old controversies, current challenges, and new directions. *Educational Psychologist*, 46, 26–47. <http://dx.doi.org/10.1080/00461520.2011.538646>
- Senko, C., & Miles, K. M. (2008). Pursuing their own learning agenda: How mastery-oriented students jeopardize their class performance. *Contemporary Educational Psychology*, 33, 561–583. <http://dx.doi.org/10.1016/j.cedpsych.2007.12.001>
- Senko, C., & Tropiano, K. L. (2016). Comparing three models of achievement goals: Goal orientations, goal standards, and goal complexes. *Journal of Educational Psychology*, 108, 1178–1192. <http://dx.doi.org/10.1037/edu0000114>
- Sheldon, K. M. (2004). *Optimal human being: An integrated multi-level perspective*. Mahwah, NJ: Erlbaum.
- Sheldon, K. M., & Elliot, A. J. (1998). Not all personal goals are personal: Comparing autonomous and controlled reasons for goals as predictors of effort and attainment. *Personality and Social Psychology Bulletin*, 24, 546–557. <http://dx.doi.org/10.1177/0146167298245010>



- Sheldon, K. M., Ryan, R. M., Deci, E. L., & Kasser, T. (2004). The independent effects of goal contents and motives on well-being: It's both what you pursue and why you pursue it. *Personality and Social Psychology Bulletin*, 30, 475–486. <http://dx.doi.org/10.1177/0146167203261883>
- Sideridis, G. D., & Kaplan, A. (2011). Achievement goals and persistence across tasks: The roles of failure and success. *Journal of Experimental Education*, 79, 429–451. <http://dx.doi.org/10.1080/00220973.2010.539634>
- Skaalvik, E. M., & Skaalvik, S. (2013). School goal structure: Associations with students' perceptions of their teachers as emotionally supportive, academic self-concept, intrinsic motivation, effort, and help seeking behavior. *International Journal of Educational Research*, 61, 5–14. <http://dx.doi.org/10.1016/j.ijer.2013.03.007>
- Smeding, A., Dompnier, B., Meier, E., Darnon, C., Baumberger, B., & Butera, F. (2015). The motivation to learn as a self-presentation tool among Swiss high school students: The moderating role of mastery goals' perceived social value on learning. *Learning and Individual Differences*, 43, 204–210. <http://dx.doi.org/10.1016/j.lindif.2015.08.036>
- Spray, C. M., John Wang, C. K., Biddle, S. J., & Chatzisarantis, N. L. (2006). Understanding motivation in sport: An experimental test of achievement goal and self-determination theories. *European Journal of Sport Science*, 6, 43–51. <http://dx.doi.org/10.1080/17461390500422879>
- Srivastava, A., Locke, E. A., & Bartol, K. M. (2001). Money and subjective well-being: It's not the money, it's the motives. *Journal of Personality and Social Psychology*, 80, 959–971. <http://dx.doi.org/10.1037/0022-3514.80.6.959>
- Standage, M., Duda, J. L., & Ntoumanis, N. (2005). A test of self-determination theory in school physical education. *The British Journal of Educational Psychology*, 75, 411–433. <http://dx.doi.org/10.1348/000709904X22359>
- Urdu, T. C. (2004a). Predictors of academic self-handicapping and achievement: Examining achievement goals, classroom goal structures, and culture. *Journal of Educational Psychology*, 96, 251–264. <http://dx.doi.org/10.1037/0022-0663.96.2.251>
- Urdu, T. C. (2004b). Using multiple methods to assess students' perceptions of classroom goal structures. *European Psychologist*, 9, 222–231. <http://dx.doi.org/10.1027/1016-9040.9.4.222>
- Urdu, T., & Mestas, M. (2006). The goals behind performance goals. *Journal of Educational Psychology*, 98, 354–365. <http://dx.doi.org/10.1037/0022-0663.98.2.354>
- Vallerand, R. J., Fortier, M. S., & Guay, F. (1997). Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social Psychology*, 72, 1161–1176. <http://dx.doi.org/10.1037/0022-3514.72.5.1161>
- Vansteenkiste, M., Lens, W., & Deci, E. L. (2006). Intrinsic versus extrinsic goal contents in self-determination theory: Another look at the quality of academic motivation. *Educational Psychologist*, 41, 19–31. [http://dx.doi.org/10.1207/s15326985ep4101\\_4](http://dx.doi.org/10.1207/s15326985ep4101_4)
- Vansteenkiste, M., Lens, W., Elliot, A. J., Soenens, B., & Mouratidis, A. (2014). Moving the achievement goal approach one step forward: Toward a systematic examination of the autonomous and controlled reasons underlying achievement goals. *Educational Psychologist*, 49, 153–174. <http://dx.doi.org/10.1080/00461520.2014.928598>
- Vansteenkiste, M., & Mouratidis, A. (2016). Emerging trends and future directions for the field of motivation psychology: A special issue in honor of Prof. Dr. Willy Lens. *Psychologica Belgica*, 56, 118–142. <http://dx.doi.org/10.5334/pb.354>
- Vansteenkiste, M., Mouratidis, A., & Lens, W. (2010). Detaching reasons from aims: Fair play and well-being in soccer as a function of pursuing performance-approach goals for autonomous or controlling reasons. *Journal of Sport & Exercise Psychology*, 32, 217–242. <http://dx.doi.org/10.1123/jsep.32.2.217>
- Vansteenkiste, M., Mouratidis, A., Van Riet, T., & Lens, W. (2014). Examining correlates of game-to-game variation in volleyball players' achievement goal pursuit and underlying autonomous and controlling reasons. *Journal of Sport & Exercise Psychology*, 36, 131–145. <http://dx.doi.org/10.1123/jsep.2012-0271>
- Vansteenkiste, M., Smeets, S., Soenens, B., Lens, W., Matos, L., & Deci, E. L. (2010). Autonomous and controlled regulation of performance-approach goals: Their relations to perfectionism and educational outcomes. *Motivation and Emotion*, 34, 333–353. <http://dx.doi.org/10.1007/s11031-010-9188-3>
- Vansteenkiste, M., Zhou, M., Lens, W., & Soenens, B. (2005). Experiences of autonomy and control among Chinese learners: Vitalizing or immobilizing? *Journal of Educational Psychology*, 97, 468–483. <http://dx.doi.org/10.1037/0022-0663.97.3.468>
- Van Yperen, N. W., Blaga, M., & Postmes, T. (2014). A meta-analysis of self-reported achievement goals and nonself-report performance across three achievement domains (work, sports, and education). *PLoS ONE*, 9, e93594. <http://dx.doi.org/10.1371/journal.pone.0093594>
- Van Yperen, N. W., Blaga, M., & Postmes, T. (2015). A meta-analysis of the impact of situationally induced achievement goals on task performance. *Human Performance*, 28, 165–182. <http://dx.doi.org/10.1080/08959285.2015.1006772>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54, 1063–1070. <http://dx.doi.org/10.1037/0022-3514.54.6.1063>
- Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology*, 96, 236–250. <http://dx.doi.org/10.1037/0022-0663.96.2.236>
- Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology*, 40, 424–431. <http://dx.doi.org/10.1016/j.jesp.2003.10.001>
- Zan, G., Xiang, P., Louis, H., Jianmin, G., & YunPeng, R. (2008). A cross-cultural analysis of achievement goals and self-efficacy between American and Chinese college students in physical education. *International Journal of Sport Psychology*, 39, 312–328.
- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81, 329–339. <http://dx.doi.org/10.1037/0022-0663.81.3.329>

(Appendix follows)

## Appendix

### Achievement Goal Questionnaire, Autonomous and Controlled Reasons Scale, and Autonomous and Controlled Achievement Goal Complex Scale (Study 4)

The first scale contains mastery goal (MAp) and performance approach goal (PAp) items, the second scale contains autonomous reasons (AR) and controlled reasons (CR) items, and the third scale represents autonomous mastery goal complex (MAp  $\times$  AR), controlled mastery goal complex (MAp  $\times$  CR), autonomous performance goal complex (PAp  $\times$  AR), and controlled performance goal complex (PAp  $\times$  CR) items.

Below you will find statements that **represent descriptions of how you might pursue goals in your classes at the university**. Please indicate how true each statement is for you.

My aim is to completely master the material presented in my classes (MAp).

My goal is to perform better than the other students. (PAp)

My goal is to learn as much as possible. (MAp)

My aim is to perform well relative to other students. (PAp)

Below you will find statements that represent **explanations for why you might pursue goals in your classes at the university**. Please indicate how true each statement is for you.

In my classes, I pursue goals because I find them highly stimulating and challenging. (AR)

In my classes, I pursue goals because I find them personally valuable goals. (AR)

In my classes, I pursue goals because I would feel bad, guilty, or anxious if I didn't do it. (CR)

In my classes, I pursue goals because I can only be proud of myself if I do so. (CR)

In my classes, I pursue goals because I have to comply with the demands of others such as parents, friends, and teachers. (CR)

In my classes, I pursue goals because others will reward me only if I achieve these goals. (CR)

Below you will find statements that represent **descriptions of how you might pursue goals** in your classes at university, together with **explanations for why you might pursue them**. Please *read each statement carefully*, and indicate how true each of it is for you.

My goal is to learn as much as possible because I find this a highly stimulating and challenging goal. (MAp  $\times$  AR)

My aim is to completely master the material presented in my classes because I would feel bad, guilty, or anxious if I didn't do it. (MAp  $\times$  CR)

My goal is to learn as much as possible because I can only be proud of myself if I do so. (MAp  $\times$  CR)

My aim is to completely master the material presented in my classes because I find this a personally valuable goal. (MAp  $\times$  AR)

My goal is to learn as much as possible because I have to comply with the demands of others such as parents, friends, and teachers. (MAp  $\times$  CR)

My aim is to completely master the material presented in my classes because others will reward me only if I achieve this goal. (MAp  $\times$  CR)

My aim is to completely master the material presented in my classes because I find this a highly stimulating and challenging goal. (MAp  $\times$  AR)

My goal is to learn as much as possible because I would feel bad, guilty, or anxious if I didn't do it. (MAp  $\times$  CR)

My aim is to completely master the material presented in my classes because I can only be proud of myself if I do so. (MAp  $\times$  CR)

My goal is to learn as much as possible because I find this a personally valuable goal. (MAp  $\times$  AR)

My aim is to completely master the material presented in my classes because I have to comply with the demands of others such as parents, friends, and teachers. (MAp  $\times$  CR)

My goal is to learn as much as possible because others will reward me only if I achieve this goal. (MAp  $\times$  CR)

My goal is to perform better than the other students because I find this a highly stimulating and challenging goal. (PAp  $\times$  AR)

My aim is to perform well relative to other students because I would feel bad, guilty, or anxious if I didn't do it. (PAp  $\times$  CR)

My goal is to perform better than the other students because I can only be proud of myself if I do so. (PAp  $\times$  CR)

My aim is to perform well relative to other students because I find this a personally valuable goal. (PAp  $\times$  AR)

My goal is to perform better than the other students because I have to comply with the demands of others such as parents, friends, and teachers. (PAp  $\times$  CR)

My aim is to perform well relative to other students because others will reward me only if I achieve this goal. (PAp  $\times$  CR)

My aim is to perform well relative to other students because I find this a highly stimulating and challenging goal. (PAp  $\times$  AR)

My goal is to perform better than the other students because I would feel bad, guilty, or anxious if I didn't do it. (PAp  $\times$  CR)

My aim is to perform well relative to other students because I can only be proud of myself if I do so. (PAp  $\times$  CR)

My goal is to perform better than the other students because I find this a personally valuable goal. (PAp  $\times$  AR)

My aim is to perform well relative to other students because I have to comply with the demands of others such as parents, friends, and teachers. (PAp  $\times$  CR)

My goal is to perform better than the other students because others will reward me only if I achieve this goal. (PAp  $\times$  CR)

Received September 6, 2016

Revision received January 10, 2017

Accepted February 13, 2017 ■



# Identifying Pre–High School Students’ Science Class Motivation Profiles to Increase Their Science Identification and Persistence

Jessica R. Chittum  
East Carolina University

Brett D. Jones  
Virginia Tech

One purpose of this study was to determine whether patterns existed in pre–high school students’ motivation-related perceptions of their science classes. Another purpose was to examine the extent to which these patterns were related to their science identification, gender, grade level, class effort, and intentions to persist in science. We collected data from pre–high school students (Grades 5 through 7, 52.5% female, and 90.7% who self-identified as White) from 2 rural public schools in Southwest Virginia. Our analysis included data from 937 questionnaires that measured students’ perceptions of empowerment/autonomy, usefulness/utility value, expectancy for success, situational interest, and caring in science class. Using cluster analysis, we identified 5 clusters (i.e., “motivation profiles”) of students: (a) low motivation, (b) low usefulness and interest but high success and caring, (c) somewhat high motivation, (d) somewhat high motivation and high success and caring, and (e) high motivation. We tested the cluster stability by cluster analyzing subsamples by year of data collection and by grade level. Significant relationships existed between these motivation profiles and students’ science identification, gender, grade level, science class effort, and intentions to persist in science. These findings may support science educators in targeting students with similar motivation profiles rather than adhering to the difficult and often unrealistic task of catering to each student’s complex needs, individually.

**Keywords:** cluster analysis, motivation, motivation profiles, person-centered research, science education

The overall aim of this study was to better understand how pre–high school (i.e., grades five through seven) students’ perceptions of their science classes can affect their science identification and intentions to persist in science-related fields. Persistence in science is important because finding well-educated and trained professionals to fill science, technology, engineering, and mathematics (STEM) positions is a national concern in the United States (Smith, 2012), in part because research and funding in STEM fields is integral to US prosperity (National Academy of Sciences [NAS], 2007). Providing educational experiences that support students’ success in science and mathematics is critical to ensuring that they are adequately prepared for STEM professions (NAS, 2007; NGSS Lead States, 2013; President’s Council of Advisors on Science and Technology [PCAST], 2012; Smith, 2012). Unfortunately, students are increasingly entering college underprepared for and uninterested in pursuing STEM fields (Osborne,

Simon, & Collins, 2003; PCAST, 2012), which has led researchers to study factors that can affect students’ motivation and intention to persist in STEM disciplines (Renninger, Nieswandt, & Hidi, 2015).

Researchers have documented that students’ motivation in science tends to wane with age (Osborne et al., 2003; Simpson & Oliver, 1990). For students’ long-term persistence in STEM fields, it is especially important to nurture their motivation and interest in science prior to eighth-grade, particularly during the pre–high school years (Maltese & Tai, 2010; PCAST, 2010, 2012; Tai, Liu, Maltese, & Fan, 2006). The pre–high school years are also critical because students who intend to persist in the sciences typically begin their formal preparation during that time (NAS, 2007). Fortunately, school climate and science teaching methods during the pre–high school years can positively impact students’ science motivation and persistence, which can help to prevent the declines that have been found in less supportive environments (Chittum, Jones, Akalin, & Schram, under review; Fortus & Vedder-Weiss, 2014; Vedder-Weiss & Fortus, 2011).

Given these findings, we were interested in how pre–high school students’ perceptions of their science classes were related to their *science identification* (i.e., the extent to which a student values science as an important part of his or her “self”; Jones, Ruff, & Osborne, 2015), because when students identify with a subject, they are more likely to persist in the subject in the future (Osborne & Jones, 2011). Therefore, identifying students’ perceptions of science class that affect their science identification may lead to strategies that teachers can use to foster students’ science identification and increase their persistence in science over time. Theoretical and empirical findings (e.g., Hidi, Renninger, &

---

This article was published Online First April 20, 2017.

Jessica R. Chittum, Department of Elementary Education and Middle Grades Education, East Carolina University; Brett D. Jones, School of Education, Virginia Tech.

This research was supported by the National Science Foundation (NSF) under Grant DRL 1029756. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

Correspondence concerning this article should be addressed to Jessica R. Chittum, Department of Elementary Education and Middle Grades Education, East Carolina University, Greenville, NC 27858. E-mail: chittumj15@ecu.edu

Nieswandt, 2015; Jones, Ruff, et al., 2015; Osborne & Jones, 2011) indicate that teaching strategies that support students’ science identification include those that are consistent with the five components of the MUSIC® Model of Motivation (MUSIC model; Jones, 2009, 2015): *eM*powerment, *U*sefulness, *S*uccess, *I*nterest, and *C*aring (MUSIC is an acronym for the first sounds of these words). Consequently, we chose to focus on students’ MUSIC model perceptions of their science classes.

We were particularly interested in examining whether patterns existed in students’ science class MUSIC perceptions. For example, students may feel empowered in their science class (high empowerment), understand the usefulness of their work in that science class (high usefulness), and feel that they can be successful in that science class (high success). Yet, they may not believe that the science classwork is interesting (low interest) or that their teacher cares about their learning (low caring). If several students have a similar pattern of these five science class perceptions, it may be possible for teachers to motivate these students by tailoring their instructional strategies to this pattern. In this study, we used cluster analysis, which is a person-centered research approach (Bergman, 2001), to determine whether these types of patterns exist for students in science classes; and, if they do, to identify the number and type of patterns.

Another purpose of this study was to investigate how patterns in students’ science class perceptions relate to their gender, grade level, class effort, and intentions to persist in science. Previous research suggests that gender can be an important factor in students’ motivation and persistence in STEM fields, with female students often less likely to be motivated and persist (Eccles, 2007; Maltese & Harsh, 2015). Moreover, older students are often less motivated than younger students (Eccles, Wigfield, Harold, & Blumenfeld, 1993; Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002). Finally, we examined students’ perceived effort in class and their reported intentions to persist (i.e., science course intentions and science career goals) because research suggests that more motivated students often put forth more effort and tend to persist in related tasks or domains (Deci & Ryan, 2000; Hidi & Renninger, 2006; Wigfield & Eccles, 2000).

Conceptual Framework

Science Identification

Students value some academic subjects more than others, and their values for these subjects can change over time (Simpkins, Davis-Kean, & Eccles, 2006). The extent to which a student values a subject as an important part of his or her “self” is defined as *domain identification* (Jones, Ruff, et al., 2015; Osborne & Jones, 2011). A domain can refer to a broader category (e.g., academics or athletics) or a narrower category (e.g., science or mathematics). Domain identification is important because it is associated with several positive outcomes, such as classroom participation and achievement (Voelkl, 1997), deep cognitive processing of course material and self-regulation (Osborne & Rausch, 2001), grade point average and academic honors (Osborne, 1997), decreased behavioral referrals and absenteeism (Osborne & Rausch, 2001), and career goals (Jones, Osborne, Paretti, & Matusovich, 2014; Jones, Paretti, Hein, & Knott, 2010; Jones, Tendhar, & Paretti, 2016).

Figure 1 shows how the variables in the present study fit into the domain identification model presented by Osborne and Jones (2011). The left side of the figure shows the social and academic background factors that can affect students’ science identification, including their science class perceptions of empowerment, usefulness, success, interest, and caring. The other parts of the figure show that science identification affects and is affected by students’ science career goals and science course intentions. These factors then affect students’ science class effort and science outcomes (e.g., grades, achievement), which then cycle back and affect the other variables in the model. Studies using structural equation modeling have confirmed the relationships of several aspects of the model in the domain of engineering (Jones, Osborne, et al., 2014; Jones, Tendhar, & Paretti, 2016). Furthermore, Jones, Ruff, et al. (2015) cited evidence from studies conducted with students in science and mathematics to demonstrate connections between students’ class perceptions of the MUSIC components and their identification with science and mathematics.

It is important to note that the MUSIC model focuses on students’ perceptions within a specific learning environment, such as a science class or a specific learning task. In contrast, domain identification focuses on students’ identification at a broader domain level, such as science.

The MUSIC Model of Motivation

Based on an extensive examination of motivation-related research, Jones (2009, 2015, 2016a) developed the multidimensional MUSIC® Model of Motivation to help teachers identify and implement teaching strategies consistent with current motivation research. The MUSIC model helps to fill a need for integrative models of motivation (Vansteenkiste & Mouratidis, 2016; Wentzel & Wigfield, 2009). The model organizes motivation-related instructional strategies into five broad categories and includes strategies that: (a) *empower* students by giving them some control over their environment, (b) demonstrate how the topic is *useful* to students’ personal goals, (c) help students believe that they can *succeed*, (d) trigger and maintain students’ situational *interest* in the topic, and (e) foster a sense of *caring* and belonging.

The MUSIC model has also been used to guide the assessment of students’ motivation-related perceptions associated with a par-

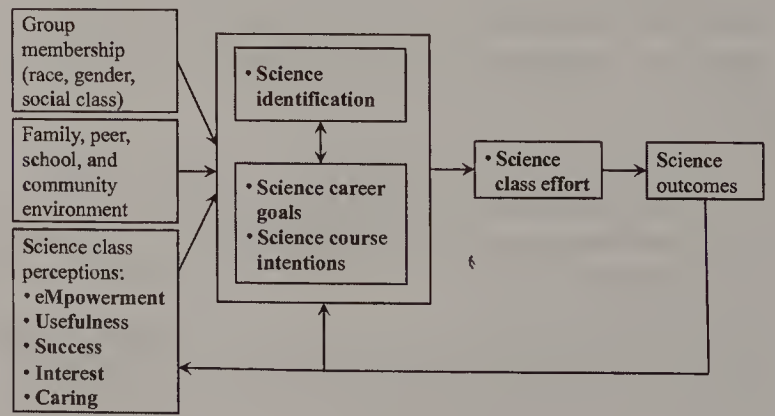


Figure 1. Relationships among the variables. Variables measured in this study are bulleted and bolded. From “Overview of the MUSIC® Model of Motivation” by B. D. Jones, 2017, p. 7. Copyright 2017 by Brett D. Jones. Reprinted with permission.



ticular class to examine the effectiveness of instructional approaches (e.g., Chittum, McConnell, & Sible, in press; Jones, Chittum, et al., 2015; Jones, Epler, Mokri, Bryant, & Parette, 2013; Jones, Ruff, Snyder, Petrich, & Koonce, 2012; Jones, Watson, Rakes, & Akalin, 2012; Lee, Kajfez, & Matusovich, 2013; McGinley & Jones, 2014) and the relationships between these perceptions and outcomes, such as domain identification, course effort, course ratings, and career goals (e.g., Jones, 2010; Jones, Osborne, et al., 2014; Jones, Tendhar, & Parette, 2016). Researchers have documented that students' class perceptions of the five MUSIC model components are distinct, yet correlated, in samples of elementary students (Jones & Sigmon, 2016), middle and high school students (Parkes, Jones, & Wilkins, 2015), and college students (Jones & Skaggs, 2016; Jones & Wilkins, 2013). Similar patterns have been shown to exist across cultures and countries (Jones, Li, & Cruz, 2017; Mohamed, Soliman, & Jones, 2013; Mora, Anorbe-Diaz, Gonzalez-Marrero, Martin-Gutierrez, & Jones, in press; Schram & Jones, 2016).

We chose to examine students' science class MUSIC perceptions in our study because they have been associated with domain identification (Jones, Osborne, et al., 2014; Jones, Ruff, et al., 2015; Jones, Tendhar, & Parette, 2016; Osborne & Jones, 2011) and for a few other reasons. First, the MUSIC model components relate to well-known constructs that have been studied over several decades and have been shown to be associated with students' class motivation and engagement (as explained in subsequent sections). Second, these constructs have been shown to be changeable by an instructor in a learning environment (Reeve, Jang, Carrell, Jeon, & Barch, 2004; Turner, Christensen, Kackar-Cam, Trucano, & Fulmer, 2014; Wang & Eccles, 2013), which is important because the constructs and associated instructional strategies would not otherwise be useful to instructors who want to increase students' motivation and engagement in a class. Third, we wanted to assess enough constructs that would allow us to explain an adequate amount of variance in educational outcomes, but not too many constructs that would result in the inclusion of constructs that overlapped significantly in definition (e.g., self-efficacy and expectancy for success). The components of the MUSIC model met this criterion because researchers have documented that the components are correlated, yet distinct (Jones et al., 2017; Jones & Skaggs, 2016; Jones & Wilkins, 2013; Parkes et al., 2015). In the following sections, we provide further description of each of the MUSIC model components.

**Empowerment.** The empowerment component of the MUSIC model refers to teaching strategies that provide students with the opportunity to become autonomous learners by encouraging perceptions of choice, freedom, and volition (Jones, 2009). By empowering students, instructors can meet their need for autonomy, which "encompasses people's strivings to be agentic, to feel like the 'origin' (deCharms, 1968) of their actions, and to have a voice or input in determining their own behavior" (Deci & Ryan, 1991, p. 243). Empowering students can meet students' psychological need for autonomy and is consistent with the tenets of self-determination theory (Deci & Ryan, 2000). In the domain of science, students who have been provided with autonomy have reported higher levels of intrinsic motivation (Berger & Hanze, 2009), interest experience (Tsai, Kunter, Ludtke, Trautwein, & Ryan, 2008), interest in science (Bulunuz, & Jarrett, 2015; Xu, Coats, & Davidson, 2012), and engagement (Hafen et al., 2012), all of which can contribute to students' science identification.

Teachers can empower students by giving them some control over their learning environment through offering meaningful choices (e.g., choices of topics and team members), offering opportunities for students to make decisions in the learning environment (e.g., lesson pace), and in welcoming students' opinions (Jones, 2009). In addition, it is important to communicate that students have an action choice or the ability to decide to be autonomous or to fully endorse relinquishing control to another (Reeve, Nix, & Hamm, 2003).

**Usefulness.** The usefulness component of the MUSIC model includes instructional strategies that encourage students to perceive their classwork (e.g., assignments, activities) as useful for their short- or long-term goals (Jones, 2009, 2015). The usefulness component is consistent with the utility value construct in expectancy-value theory (Eccles et al., 1983; Eccles & Wigfield, 1995). As explained by Wigfield and Eccles (2000), "Utility value or usefulness refers to how a task fits into an individual's future plans" (p. 72).

Perceptions that learning tasks are useful or instrumental in achieving academic and personal goals can positively affect domain identification (Jones, Osborne, et al., 2014; Jones, Tendhar, & Parette, 2016) and many constructs closely related to science identification, including interest (Nieswandt & Shanahan, 2008; Reynolds, Mehalik, Lovell, & Schunn, 2009), motivation (Simons, Vansteenkiste, Lens, & Lacante, 2004), persistence (De Volder & Lens, 1982; Miller, Greene, Montalvo, Ravindran, & Nichols, 1996; Simons et al., 2004), engagement (Miller et al., 1996; Simons et al., 2004), effort (De Volder & Lens, 1982; Miller et al., 1996; Simons et al., 2004), and intention to study in a specific field (Jones et al., 2010). To support students' perceptions of usefulness in an educational environment, instructors can: design tasks and activities that relate to students' long-term goals; connect content, routines, and strategies to the real world through rationales and by defining real-life implications; implement experiential, hands-on learning; and incorporate personally relevant topics (Hulleman, Durik, Schweigert, & Harackiewicz, 2008; Jones, 2009; e.g., Jones, Chittum, et al., 2015).

**Success.** The success component of the MUSIC model includes teaching strategies that support students' perceptions that they can succeed if they put forth the appropriate effort (Jones, 2009, 2015). This component is consistent with constructs such as expectancy for success (Wigfield & Eccles, 2000), competence motivation (Elliot & Dweck, 2005; White, 1959), the psychological need for competence (Deci & Ryan, 2000, 2012), and self-efficacy (Bandura, 1986).

Researchers have related high expectancies for success and competence beliefs to several constructs associated with science identification, including intentions to persist in science (e.g., pursue science careers, courses, and tasks; DeBacker & Nelson, 2000; Ireson & Hallam, 2005; Rudasill & Callahan, 2010; Simpkins et al., 2006); increased engagement (Hoffmann, 2002; Scogin & Stuessy, 2015); higher performance (Hoffmann, 2002); increased strategy use (Cheung, 2015); and positive affect for science (DeBacker & Nelson, 2000; Hoffmann, 2002). In studies examining the relationships between the MUSIC model components and domain identification (e.g., Jones, Osborne, et al., 2014; Jones, Tendhar, & Parette, 2016), researchers have often equated the success component of the MUSIC model to the expectancy for success construct (Eccles et al., 1983), which has been shown to



contribute to students' level of domain identification. These studies have measured expectancy for success (as opposed to self-efficacy, competence, or self-concept) in part because students' MUSIC perceptions have been assessed at the class level as opposed to the task level, which would be appropriate for measuring the self-efficacy construct (Bong & Skaalvik, 2003). Furthermore, students' ratings of expectancy for success have not been shown to be empirically distinct from their ratings of ability, competence, and self-concept (Eccles & Wigfield, 1995; Eccles et al., 1993); therefore, it is redundant to assess more than one of these constructs.

Teachers can support students' success perceptions in a variety of ways, such as by providing: attainably challenging tasks and learning goals; clear and realistic expectations; meaningful, timely, and constructive feedback that can be implemented and is applicable to future learning; opportunities for success if students put forth effort; and opportunities to practice and master concepts (Jones, 2009). Teachers can also foster malleable beliefs about intelligence, teach lessons considering novice versus expert understandings, and break difficult tasks into attainable chunks to nurture positive ability perceptions (Jones, 2009, 2015).

**Interest.** The interest component of the MUSIC model pertains to instructional strategies that stimulate interest in the academic activity, content, or domain (Jones, 2009). Interest can be defined as "liking and willful engagement in a cognitive activity" (Schraw & Lehman, 2001, p. 23); therefore, it includes both an affective component of positive emotion and a cognitive component of concentration (Hidi & Renninger, 2006). Students' interests can progress along a continuum in which triggered *situational interest* (which is short-term and context-specific) can lead to well-developed *individual interest* (which is more enduring than situational interest; Hidi & Renninger, 2006). Because the intent of the present study was to investigate students' perceptions of their current science class, we focused on their situational interest rather than their longer-term individual interests. Our rationale was that, regardless of students' level of individual interest in science, instructors can strive to design instruction that is situationally interesting to students. In addition, situational interest is a necessary condition for the development of individual interest (Hidi & Renninger, 2006), which is similar in many ways to domain identification (see Jones, Ruff, et al., 2015 for a discussion). Thus, if teachers can trigger and maintain students' situational interest, they may be able to develop students' individual interest and identification in the domain.

Situational interest is consistent with constructs such as intrinsic motivation (Deci, 1975), intrinsic interest value (Eccles & Wigfield, 1995), and flow (Csikszentmihalyi, 1990), and can influence a variety of factors, including engagement, attention, persistence, goals, strategy use, enjoyment, and performance (Hidi & Harackiewicz, 2000; Hidi & Renninger, 2006; Schraw & Lehman, 2001). In science, situational interest has been associated with continued engagement in science activities (Spiegel, McQuillan, Halpin, Matuk, & Diamond, 2013) and more motivation to learn science (Barak, Ashkar, & Dori, 2011; Rosen, 2009). Teachers can stimulate situational interest in many ways, such as by inciting curiosity and/or strong emotions, introducing novelty, using a variety of instructional tools and/or tasks, facilitating social interaction, connecting content to background knowledge and prior experiences, and using humor (Bergin, 1999; Hidi et al., 2015).

**Caring.** The caring component of the MUSIC model includes instructional strategies aimed to help students believe that their instructors and classmates care about their learning and general well-being (Jones, 2009, 2015). The caring component of the MUSIC model is consistent with constructs such as caring (Noddings, 1992), belonging (Baumeister & Leary, 1995), relatedness (Deci & Ryan, 2000, 2012), and attachment (Ainsworth, 1973; Bowlby, 1969). Positive interactions with instructors and peers can positively influence motivation-related outcomes (Wentzel, 1997; Wentzel, Battle, Russell, & Looney, 2010). Furthermore, when students have healthy, secure attachments with teachers, parents, and peers, they are more likely to experience an increase in academic performance, academic motivation, emotional development, and social skill development (Bergin & Bergin, 2009). Specifically in the domain of science, when students perceive care, support, and/or positive relations with others in the learning environment, they are more likely to hold positive attitudes about and values for science (Jen, Lee, Chien, Hsu, & Chen, 2013), to develop their science identity (Lee, 2002; Stake & Nickens, 2005), and to intend to persist in science (Jacobs, Finken, Griffin, & Wright, 1998; Stake & Nickens, 2005).

Instructors can encourage positive perceptions of caring and feelings of belonging through their classroom interactions (Jones, 2009). In Wentzel's (1997) study, students described caring instructors as those who emphasized a democratic style, respected the individuality of students, provided positive and meaningful feedback, and went the "extra mile" in teaching and planning. Caring can also be nurtured by supporting students' educational goals; demonstrating care and concern for achieving learning objectives, personal goals, and well-being; carefully designing instruction to encourage student learning; providing opportunities for positive interactions with peers; and making oneself available for academic support after hours (Jones, 2009, 2015).

**Gender differences.** Male and female students have been shown to differ on their perceptions of some of the MUSIC model components (Meece, Glienke, & Burg, 2006). For example, females tend to have lower expectancies for success about science-related proclivities (Bong, Lee, & Woo, 2015). In another study, females perceived high school physical science courses to be less useful than did the male students, which led them to enroll in fewer physical science courses (Eccles, 2007). Also, classroom experiences appear to have more of an effect on female students' interests in science than for males. For example, in one study male students were more likely to report that their interest was triggered from building or tinkering with mechanical objects or electronics, or reading books and magazines (Maltese & Harsh, 2015). These findings suggest that teachers may be able to play an especially important role in helping female students improve their perceptions of success, interest, and usefulness in science.

## Person-Centered Approaches

Although variable-centered approaches to data analysis have been effective at identifying trends in students' interests and motivation-related perceptions in science over time (Simpson & Oliver, 1990; Simpkins et al., 2006), these approaches do not allow teachers to understand how students' motivation perceptions interact with one another during a class to motivate students. Variable-centered approaches investigate the effects of isolated



variables linearly, rather than “motivational phenomena as continuously emerging systems of dynamically interrelated components” (Kaplan, Katz, & Flum, 2014, para. 4) that are multidimensional, complex, and context-bound (Turner & Meyer, 2000). A different approach to data analysis, often named a *person-centered* or *person* approach (e.g., cluster analysis), allows researchers to focus on the complex nature of the individual, and motivation constructs are studied as dynamic, interactionistic processes (Bergman, 2001; Kaplan, Katz, & Flum, 2012; Vansteenkiste & Mouratidis, 2016). Hence, in person-centered approaches, the variable is not central; rather, the person is central because the dynamic interplay of multiple variables is studied by investigating patterns and relationships among them (Bergman, 2001). Consequently, in cluster analysis, for example, researchers can examine a more integrated profile of the individual in which those with similar patterns of relationships among variables are organized into a cluster or “profile” (Bergman, 2001; Wormington, Corpus, & Anderson, 2012).

We used cluster analysis in the present study to identify patterns in students’ MUSIC perceptions of their science classes. Several recent studies have used cluster analysis to examine students’ profiles at the school or academic level (Bowers & Sprott, 2012; Hayenga & Corpus, 2010; Meece & Holt, 1993; Ratelle, Guay, Vallerand, Larose, & Sénécal, 2007; Schwinger, Steinmayr, & Spinath, 2012; Tuominen-Soini, Salmela-Aro, & Niemivirta, 2011; Vansteenkiste, Sierens, Soenens, Luyckx, & Lens, 2009; Wormington et al., 2012), at the domain level (Chen, 2012; Conley, 2012; Hartwell & Kaplan, 2014; Turner, Thorpe, & Meyer, 1998), at the class level (Daniels et al., 2008; Shell & Husman, 2008; Shell & Soh, 2013), or at the task level (Geiser, Lehmann, & Eid, 2006; Janssen & Geiser, 2010). However, no studies have focused on pre-high school students in science classes, which is the population of interest in the present study.

### Research Questions

To address the following research questions, we surveyed fifth-, sixth-, and seventh-grade students about their science class perceptions, science identification, science class effort, and intentions to persist in science. RQ1: Can students’ science class perceptions be used to categorize students into groups with similar motivation profiles? RQ2: If different profiles can be identified, are students’ class-related motivation profiles associated with their science identification, science class effort, and intentions to persist in science? RQ3: If different profiles can be identified, does membership in the profiles vary by students’ gender or grade level? We expected that, at the minimum, a high motivation profile and a low motivation profile would emerge. Our hypothesized “high” motivation profile would include students who rated all of their science class MUSIC perceptions highly. The “low” motivation profile would include students who rated all of their class MUSIC perceptions low. To maintain the exploratory nature of this investigation and to avoid preconceptions that could influence our analysis, we did not hypothesize further about the profiles. A primary limitation of cluster analysis is that researcher judgment may unduly influence the cluster solution (Burns & Burns, 2008); thus, we intentionally avoided making specific hypotheses regarding the cluster solution so as to approach the analysis as objectively as possible.

We further hypothesized that students in a high motivation profile would put forth more effort in science class and report greater intentions to persist in science than students in a low motivation profile. In addition, we hypothesized that students in higher grade levels would be in lower motivation profiles due to commonly reported declines in motivation over time (Eccles et al., 1993; Jacobs et al., 2002), and that there would be fewer female students in higher motivation profiles because of the gender gap associated with STEM fields (Meece et al., 2006).

The results of this study may be used to help educators target students with similar class-related motivation profiles, rather than adhere to the difficult and often unrealistic task of catering to each student’s individual complex needs. Moreover, it may be possible to identify students with class-related motivation profiles that are more or less likely to pursue science-related majors or careers. Using profiles, teachers could more intentionally target students’ motivation in science classrooms and increase the likelihood that more students will engage in science, either by choosing a science-related career or by becoming a more scientifically literate member of society.

## Method

### Participants

The participants were students in grades five, six, and seven from two rural public schools in Southwest Virginia. We collected data at three time-points and received responses from 323 students in 2012 (84% of all students in those grades at the schools), 320 students in 2013 (87% of all possible students), and 291 students in 2014 (76% of all possible students). This sample included a total of 934 completed questionnaires (some students completed a questionnaire for two or three years), with 398 completed questionnaires (178 students) representing students assessed at multiple time points and 536 students assessed only once. Hereafter, we refer to each completed questionnaire as one “case.”

The majority (90.7%) of the students identified as White and the others identified as Black or African American (1.6%), Hispanic (1.1%), Asian or Pacific Islander (1.1%), American Indian (2.6%), or “other” (2.7%), and two students chose not to answer. Slightly over half of the students (52.5%) were female. According to state guidelines, both schools were considered to comprise a high proportion of low-income students and qualified for federal Title I funds (Virginia Department of Education [VDOE], 2012; VDOE Office of School Nutrition Programs, 2014). The science curriculum in the fifth- and seventh-grades comprised earth/space, life, and physical science content. In sixth-grade, the science curriculum comprised both earth/space science and physical science. Each school contributed approximately half of the cases (47.7% and 52.3%).

### Procedures

In May of 2012, 2013, and 2014, all fifth-, sixth-, and seventh-grade students present at two K-7 schools in the same county completed a questionnaire related to their perceptions about science. Students had been enrolled in their science classes since the beginning of the school year and the questionnaires were administered near the end of each school year. The schools had seven



science teachers in the three grade levels, with one at each grade level (except for fifth-grade at one school, which added a second science teacher in 2014). We obtained Institutional Review Board approval prior to conducting the study.

Measures

The questionnaire was titled generically as a “Science Questionnaire” and was part of a larger study that examined students’ motivation-related perceptions about their current science classes, their motivation beliefs about science, and their demographic information. The items were scaled using a 6-point Likert-type format with the following descriptors: 1 = *strongly disagree*, 2 = *disagree*, 3 = *mostly disagree*, 4 = *mostly agree*, 5 = *agree*, and 6 = *strongly agree*.

**Science class perceptions.** To develop profiles of students’ perceptions of their science class, we measured each of the five components of the MUSIC model using the MUSIC® Model of Academic Motivation Inventory (MUSIC Inventory; Jones, 2016b). We used the middle/high school version of the MUSIC Inventory (Jones, 2016b) that was designed to measure middle and high school students’ science class perceptions using the five MUSIC model components. Table 1 shows the MUSIC model components, their definitions, and the related constructs in the MUSIC Inventory (Jones, 2016b). Although the MUSIC Inventory was designed specifically to measure students’ class perceptions of the five MUSIC model components, it is noteworthy that (a) it is possible that the inventory does not measure the range of possible perceptions within each MUSIC component, and (b) other instruments that measure the MUSIC model components might focus on different aspects of the components. Nonetheless, the constructs measured with the MUSIC Inventory have been shown to separate into distinct constructs using factor analysis (Jones & Skaggs, 2016; Parkes et al., 2015; Schram & Jones, 2016).

Table 2 includes example items from each MUSIC Inventory scale. Cronbach’s alpha values for the MUSIC Inventory have been shown to be acceptable for fifth- to twelfth-grade students in music and band ensemble classes (Parkes et al., 2015; empowerment  $\alpha = .73$ , usefulness  $\alpha = .86$ , success  $\alpha = .92$ , interest  $\alpha = .91$ , caring  $\alpha = .92$ ) and for the students in the present study (see Table 3). For this study, we also used LISREL 8.8 to conduct three confirmatory factor analyses (CFAs; one for each of the three years) and included 22 items: the 18 items from the five MUSIC Inventory scales and the four items from the science identification scale. The fit indices we computed—the Comparative Fit Index (CFI), the Standardized Root Mean Square Residual (SRMR), and

the Root Mean Square Error of Approximation (RMSEA)—were all within acceptable limits (see Table 3; Browne & Cudeck, 1993; Byrne, 2001; Hu & Bentler, 1999; Kline, 2005). Thus, the CFAs documented that not only were the five constructs measured by the MUSIC Inventory distinct but also that these five constructs were distinct from the science identification construct.

**Science identification.** We measured science identification using a four-item Identification with Science scale initially based on the four-item measure used by Schmader, Major, and Gramzow (2001;  $\alpha = .78$ ) and used in the domain of engineering to measure the extent to which engineering students identified with engineering (e.g.,  $\alpha = .84$  and  $0.89$  in Jones et al., 2010;  $\alpha = .92$  in Jones et al., 2014). Table 2 includes a sample item. Scores from this measure have been positively related to a variety of students’ beliefs, including career goals (Jones et al., 2010, 2014; Jones, Tendhar, & Paretti, 2016). Cronbach’s alpha values for the students in the present study are presented in Table 3.

**Science class effort.** We measured science class effort with a four-item measure used by Jones (2010) that was based on the Effort/Importance Scale that is part of the Intrinsic Motivation Inventory (Plant & Ryan, 1985). This scale measures the amount of perceived effort that students put forth in a class. See Table 2 for an example item. Cronbach’s alpha values were acceptable in the present study (2012  $\alpha = .87$ , 2013  $\alpha = .87$ , 2014  $\alpha = .85$ ) and have been shown to be acceptable in past studies ( $\alpha = .84$ ,  $.84$ ,  $.86$ , and  $.84$  in Jones, 2010;  $\alpha = .95$  in Jones et al., 2014).

**Science course intentions.** We developed one item to measure students’ desire to enroll in more science courses in the future (see Table 2). This item was based on similar items used by Hulleman et al. (2008) to measure students’ subsequent interest.

**Science career goals.** We used a two-item measure of science career goals that was based on similar single items that have been used to measure the likelihood that students’ careers would directly relate to engineering (e.g., Jones et al., 2014; Jones, Tendhar, & Paretti, 2016), which serves as a measure of intent to persist in a science-related field. A sample item is included in Table 2. These items have been associated with other motivation-related constructs in ways consistent with theories (Jones et al., 2014; Jones, Tendhar, & Paretti, 2016). In this study, we refer to both science course intentions and science career goals as measures of students’ intentions to persist in science. Cronbach’s alpha values were acceptable for the present study, 2012  $\alpha = .83$ , 2013  $\alpha = .77$ , 2014  $\alpha = .82$ .

Table 1  
The MUSIC Model Components, Definitions, and Related Constructs

MUSIC component	Definitions	
	The degree to which a student perceives that:	Related constructs
Empowerment	he or she has control of his or her learning environment	autonomy
Usefulness	the classwork is useful to his or her future	utility value
Success	he or she can succeed at the classwork	expectancy for success
Interest	the instructional methods and classwork are interesting or enjoyable	situational interest
Caring	the teacher cares about whether the student succeeds in the classwork and cares about the student’s well-being	caring

Note. Based on Jones (2016b).



Table 2  
*Example Items*

Construct	Example item	No. items
Empowerment	I have control over how I learn the content in science class.	4
Usefulness	In general, science classwork is useful to me.	3
Success	During science class, I feel that I can be successful on the classwork.	4
Interest	The science classwork is interesting to me.	3
Caring	My science teacher cares about how well I do in science class.	4
Science identification	Being good at science is an important part of who I am.	4
Science class effort	I put a lot of effort into my science class.	4
Science course intentions	I would like to take more science courses in the future.	1
Science career goals	My future career will involve science.	2

## Analysis

Using mean scores for students' science class perceptions of each MUSIC model component, we followed a two-step clustering procedure recommended by Bacher, Wenzig, and Vogler (2004) that has been used in several studies (Hartwell & Kaplan, 2014; Huberty, Jordan, & Brandt, 2005; Vansteenkiste et al., 2009; Wormington et al., 2012): (1) hierarchical agglomerative analysis (following Ward's method) followed by (2) *k*-means analysis. This process includes both hierarchical and nonhierarchical methods to find the most appropriate cluster fit, in which the second analysis serves to more effectively demarcate clusters developed in the first (Bacher et al., 2004). Using SPSS version 22 software, we first ran a hierarchical analysis to determine the optimal number of clusters and preliminary *cluster centers* (i.e., means for each MUSIC model variable in each cluster; Burns & Burns, 2008) and then *k*-means analysis for validation and to obtain the final cluster centers (i.e., means; Mooi & Sarstedt, 2011). As an ancillary purpose of this paper, we included a detailed description of our use of cluster analysis in this section because cluster analysis is not as widely used in educational research as many other methods.

Hierarchical agglomerative cluster analysis is a process through which clusters or groups form when individual cases (i.e., a single student's five-dimensional response, which includes one value for each of the five MUSIC components measured) are amalgamated at each step of the analysis until, at the final step, all cases combine into one large cluster (Bartholomew, Steele, Galbraith, & Moustaki, 2008; Kaufman & Rousseeuw, 1990). The researcher can then determine at which stage the most appropriate number of clusters formed. During the analysis, cases with similar responses are amalgamated. Initially, each case represents a single-case cluster, which generally form the initial cluster centers. Then, at each

step, every case or cluster is compared with other cases or clusters, and pairings are selected that represent the least amount of lost information (i.e., the least sum of squares, or differences from the overall cluster center; Bartholomew et al., 2008). Ward's method reduces variance within clusters by computing squared Euclidean distances, which sums the squared differences across every variable in the analysis during each stage (Norušis, 2011). By minimizing the distance measures from each case and the cluster center, cases that combine into the same cluster are more similar than those assigned to other clusters. When cases are amalgamated in this way, the foremost consideration is (the inevitable) loss of information; the focus is to minimize difference/dissimilarity measures to develop fairly homogeneous groups (Bartholomew et al., 2008; Norušis, 2011). Before computing Ward's procedure, we sorted the existing data randomly. Hierarchical analysis can be sensitive to order because the analysis begins at one case and systematically assesses difference between each case such that the order of the cases can affect the initial cluster centers and, thus, how clusters begin forming (Norušis, 2011).

To select the optimum number of clusters for our cluster solution (i.e., a stopping point in the cluster analysis) and maximize internal validity (Bacher, 2002), we implemented two methods: (a) we examined the fusion coefficients provided in an agglomeration schedule (an SPSS output) for measures of change as clusters merged; and (b) we used the Bayesian Information Criterion (BIC) to determine the optimal number of clusters and model (Fraley & Raftery, 1998; Nylund, Asparouhov, & Muthén, 2007) using the R Project for Statistical Computing (R Project) software (Mclust package).

When a large decrease in the fusion coefficient is notable between two steps, clusters merged that caused a substantial

Table 3  
*Cronbach's Alpha Values and Fit Indices*

Year	<i>n</i>	Cronbach's alpha values						CFI	SRMR	RMSEA
		M	U	S	I	C	Identity			
2012	321	.72	.78	.83	.77	.84	.82	.97	.052	.069
2013	308	.72	.83	.77	.76	.79	.83	.96	.058	.076
2014	284	.78	.82	.85	.78	.77	.83	.98	.050	.058

*Note.* CFI, RMSEA, and SRMR are values from CFAs that were conducted with all of the items from the MUSIC Inventory (i.e., empowerment [M], usefulness [U], success [S], interest [I], and caring [C]) and science identification (Identity) scales for each year separately.

change in overall within-cluster dissimilarity. Smaller change coefficients between subsequent steps indicate that those clusters bear similar heterogeneity; thus, merging clusters during those stages “adds much less to distinguishing between cases” (Burns & Burns, 2008, p. 560). We designated the stopping point (i.e., the *cluster solution*, or number of clusters) when cluster coefficients indicated a large change such that later steps became markedly more similar (Burns & Burns, 2008, p. 561; Norušis, 2011). When multiple cluster solutions appeared suitable, we examined each solution’s cluster centers and selected the solution representative of the most parsimonious and theoretically meaningful model (Bacher, 2002; Shell & Soh, 2013; Turner et al., 1998; Vansteenkiste et al., 2009). Finally, we tested other potentially viable solutions to determine whether they rendered more appropriate, meaningful solutions (Shell & Soh, 2013), analyzing the relative validity of the cluster solutions (Bacher, 2002).

Next, we used BIC for expectation-maximization to validate the previous analysis and confirm the number of clusters. In this procedure, data are partitioned through a blend of agglomerative hierarchical clustering procedures for Gaussian mixture models, and the expectation-maximization algorithm (Fraley & Raftery, 1998). Then, BIC is implemented to compare multiple models and determine the optimal solution. For this test, a parameter for modeling (i.e., maximum number of clusters/models) is set in advance. We ran the test twice: once with 100 as the parameter and a second time with 12.

A limitation of hierarchical analysis is that, once a case has been assigned a particular cluster, it cannot be unassigned (Asendorpf, Borkenau, Ostendorf, & van Aken, 2001; Kaufman & Rousseeuw, 1990). In other words, cases cannot move to different, perhaps more appropriate clusters later during the analysis as the clusters take shape and deviate naturally from the initial formation. Hence, it is important to complete an additional clustering method using the cluster solution determined with Ward’s method (Norušis, 2011). With this limitation in mind, we computed *k*-means cluster analysis as a secondary test, which is considered a validation procedure (Mooi & Sarstedt, 2011) and test of stability (Bacher, 2002). *K*-means cluster analysis involves selecting a predetermined number of clusters (*k*)—we used the number of clusters defined by Ward’s method—to “fine tune” the cluster centers (Bacher et al., 2004; Huberty et al., 2005; Wormington et al., 2012). The *k*-means procedure is used as a secondary test because hierarchical analysis is needed initially to determine the optimal number of clusters, which serves as *k*. Unlike hierarchical cluster analysis, *k*-means clustering allows cases to flow through multiple iterations such that cases can change their cluster assignment as the analysis matures; thus, the resulting cluster centers are more reliable and accurate. Iterations begin with a set of cluster centers whereby cases are classified per their distance to that centroid (Norušis, 2011). Then, each cluster center from the previous step is recomputed. Next, cases are assigned again to cluster centers based on the new averages and the aforementioned steps are repeated until there is little change in the cluster centers between steps (Norušis, 2011). The final iteration ends with each case assigned to a permanent cluster and the final cluster centers are computed (Norušis, 2011). To assess the reliability of the clusters, we compared the hierarchical and *k*-means cluster solutions using Cohen’s kappa<sup>1</sup> ( $\kappa$ ; Reilly, Wang, & Rutherford, 2005), with a

value considered acceptable at .60 or higher when comparing clusters (Asendorpf et al., 2001; Vansteenkiste et al., 2009).

We validated our cluster solution in several ways. First, we ran the same analyses with multiple subsets of the population, including analyzing the data from multiple years and grade levels. In addition, we recomputed several cluster analyses with cases sorted randomly to assess stability (Bacher, 2002). Then, we used a formal double-split cross-validation procedure in which we split the subsample into two random halves, recomputed the two-step clustering procedure followed by a nearest neighbor analysis as a reliability measure (Breckenridge, 2000). Then, we compared the nearest neighbor solution to the two-step clustering solution using Cohen’s  $\kappa$ , with  $\kappa \geq .60$  considered acceptable fit for this test (Breckenridge, 2000; Wormington et al., 2012). Only to verify that the clusters were statistically different (Mooi & Sarstedt, 2011) and produce more evidence for the internal validity (Bacher, 2002), we examined one-way ANOVAs with the five clustered variables as the dependent variables and cluster membership as the factor.

To determine appropriate cluster typology while preserving the multivariate properties of the analysis, we computed a discriminant function analysis (Burns & Burns, 2008; Jung, Owusu-Antwi, & An, 2006; Weissman & Magill, 2008). Discriminant analysis is a multivariate method that distinguishes between groups based on several variables (Galbraith & Jiaqing, 1999) and can be used to characterize, or profile, clusters (Jung et al., 2006). In essence, the variables that contributed most in distinguishing between groups are highlighted (Hale & Glassman, 1986). We included the cluster membership as the dependent variable and averages of each MU-SIC model variable as the independent variables. In the present study, discriminant analysis served in a descriptive function only; its common utility as a function of probability was irrelevant (Fraley & Raftery, 2002; Weissman & Magill, 2008).

To examine RQ2, we tested the predictive validity of the clustering solution by running several one-way ANOVAs with theoretically correlated variables, including science identification, science class effort, science course intentions, and science career goals. Finally, to address RQ3, we ran Pearson chi-square tests to investigate differences between genders and grade levels within the clusters.

## Results

Table 4 includes correlations among all tested variables from 2013 and 2014. As predicted by the domain identification model (Osborne & Jones, 2011), most of the correlations among class perceptions, science identification, science career goals, science

<sup>1</sup> Cohen’s kappa ( $\kappa$ ) is frequently used as a measure of interrater reliability (von Eye & Mun, 2005). During the test, two categorical variables are compared for each case. The total number of instances in which both categorical variables were the same for each case in a data set (i.e., both raters entered identical codes or categories for an excerpt) is computed, which is also considered the level of agreement among the raters (von Eye & Mun, 2005). An acceptable level of agreement can be judged from  $\kappa$ , which ranges in value from 0.00 to 1.00, and can depend on the purpose of the test. According to Landis and Koch (1977), greater than 0.81 is considered near perfect agreement, 0.61 to 0.80 substantial, 0.41-0.60 moderate, 0.21 to 0.40 fair, 0.01 to 0.20 slight, and 0.00 poor. Similarly, Fleiss (1981) posited that greater than 0.75 is considered excellent, 0.40 to 0.75 is considered good, and below 0.40 is considered poor.



Table 4  
Correlations and Descriptive Statistics (2013 and 2014)

Variable	1	2	3	4	5	6	7	8	9	10
1. Grade level	—	-.183**	.006	.036	-.177**	-.133*	-.123*	-.116	-.135*	-.228**
2. Sci. identification	-.265**	—	.490**	.541**	.798**	.520**	.570**	.671**	.629**	.474**
3. Sci. career goals	-.127*	.571**	—	.657**	.387**	.323**	.644**	.325**	.399**	.179**
4. Sci. course intent	-.092	.545**	.724**	—	.469**	.364**	.524**	.389**	.492**	.187**
5. Effort	-.248**	.841**	.451**	.438**	—	.521**	.566**	.692**	.670**	.510**
6. Empowerment	-.141*	.474**	.361**	.332**	.533**	—	.581**	.482**	.611**	.376**
7. Usefulness	-.159**	.657**	.678**	.602**	.600**	.517**	—	.464**	.698**	.277**
8. Success	-.188**	.648**	.417**	.424**	.675**	.505**	.497**	—	.614**	.643**
9. Interest	-.175**	.704**	.526**	.532**	.705**	.565**	.712**	.586**	—	.444**
10. Caring	-.175**	.421**	.241**	.223**	.496**	.389**	.307**	.606**	.389**	—
2014 <i>M</i> ( <i>SD</i> )	5.95 (0.88)	4.53 (1.11)	3.39 (1.57)	3.30 (1.70)	4.74 (1.07)	4.17 (1.16)	4.14 (1.33)	4.85 (1.47)	4.26 (1.27)	5.12 (0.97)
2013 <i>M</i> ( <i>SD</i> )	6.06 (0.84)	4.36 (1.17)	3.23 (1.55)	3.21 (1.77)	4.61 (1.18)	4.12 (1.12)	3.96 (1.38)	4.83 (1.01)	4.20 (1.27)	4.98 (1.08)

Note. The 2014 sample is in the upper diagonal of the matrix and the 2013 sample is in the lower diagonal of the matrix. Results are available for the 2012 sample upon request. Sci. = science; Sci. course intent = science course intentions. 2014 *n* = 284. 2013 *n* = 308.

\*  $p < .05$ . \*\*  $p < .01$ .

course intentions, and science class effort were positive and statistically significant, and most of them were moderate to strong. An exception was that caring was only weakly correlated with both science course intentions and career goals. Grade level was correlated negatively with all of the variables except for science course intentions in the 2013 and 2014 samples, and science career goals in the 2014 sample. This finding that students at the higher grades reported lower science class perceptions than students at the lower grades is consistent with previous findings (Wigfield, Eccles, Schiefele, Roeser, & Davis-Kean, 2006). The overall patterns of these correlations are consistent with theory and previous research (Eccles & Wigfield, 1995; Jones et al., 2014; Jones & Skaggs, 2016; Osborne & Jones, 2011; Wang & Eccles, 2013).

### Cluster Analyses

The first step of our analysis involved removing univariate and multivariate outliers. We removed cases with one or more variable means 3 standard deviations above or below each overall variable mean (Vijendra & Shivani, 2014), which included 18 (1.9%) univariate outliers. Next, we ran initial *k*-means cluster analyses to identify and remove cases that formed any extremely small clusters (i.e., clusters with very few cases; Jiang, Tseng, & Su, 2001; Kaufman & Rousseeuw, 1990), which accounted for three multivariate outliers. We utilized this procedure because *k*-means clustering is especially sensitive to outliers, often forming small *N* clusters that use outlier cases as cluster centers (Kaufman & Rousseeuw, 1990; Norušis, 2011). In all, we removed 21 outliers (2.2%), which left a total sample of 913 cases.

A limitation of cluster analysis is that the method may not produce any meaningful or repeatable solutions, as clustering is primarily an exploratory method and can depend heavily on the structure of the sample (Bartholomew et al., 2008; Lange, Roth, Braun, & Buhmann, 2004). A cluster solution is considered more robust and stable when it is repeated under different circumstances (e.g., different clustering algorithms, reordered cases, diverse samples or subsamples; Bacher, 2002; Lange et al., 2004; Norušis, 2011). Cluster-analyzing subsamples characterized by specific variables (e.g., grade level, year) can test whether those variables influence the cluster solutions (Bacher, 2002). Accordingly, we

computed multiple cluster analyses in two main stages to explore the profiles and examine their stability across subsamples. First, we conducted two-step cluster analyses for cases at each year point (2012 *n* = 321, 2013 *n* = 308, 2014 *n* = 284) in three separate two-step analyses, which were our primary cluster analyses. Second, to test the stability of the cluster solutions obtained in the first stage (Bacher, 2002), we computed the two-step clustering procedure for cases at each grade level (fifth *n* = 324, sixth *n* = 263, seventh *n* = 326) across the three years in three separate two-step analyses. We conducted the final tests only to confirm stability of the motivation profiles already identified.

In cluster analysis, at least two observations for each variable (2:1) with a minimum of 200 observations is considered an acceptable ratio (Egan, 1984). Because we had 913 observations and five variables (913:5), our sample of 913 cases was more than adequate. In addition, our sample sizes for the year analyses (2012 *n* = 321, 2013 *n* = 308, 2014 *n* = 284) and grade level analyses (fifth *n* = 324, sixth *n* = 263, seventh *n* = 326) were also adequate. Similar cluster analyses in academic motivation literature included comparable sample sizes (Vansteenkiste et al., 2009; Hartwell & Kaplan, 2014; Shell & Soh, 2013; with sample sizes of 291 and 484 [two analyses], 139, and 233, respectively).

**Stage I: Clusters per year.** We selected a five-cluster solution as the best description of the data for the three hierarchical analyses at each year point (2012, 2013, and 2014) based on our consideration of the fusion coefficients and our theoretical interpretation. We reached this solution for each year independently. Our secondary method using BIC also rendered a five-cluster solution under both parameters (12, 100 set as the maximum number of clusters allowed), confirming this choice. Furthermore, the cluster centers aligned between years, as shown in Table 5 and Figures 2 and 3. We also examined *k*-means analyses of three- and four-cluster solutions, and determined that they did not provide more meaningful or interpretable solutions. The four-cluster solution combined cases from Clusters 2 and 4 from the five-cluster solution. The five-cluster solution added meaning by parsing out those students whose perceived usefulness and interest were either somewhat more negative (Cluster 2) or positive (Cluster 4) and assigned them to separate clusters. The three-cluster solution,

Table 5  
Five-Cluster Solution: Comparisons Among Years

MUSIC component	Year	Clusters				
		1	2	3	4	5
Empowerment	2012	2.7 <sup>sl</sup>	4.0 <sup>sh</sup>	4.0 <sup>sh</sup>	3.8 <sup>sh</sup>	5.4 <sup>h</sup>
	2013	2.6 <sup>sl</sup>	4.2 <sup>sh</sup>	4.0 <sup>sh</sup>	4.1 <sup>sh</sup>	5.1 <sup>h</sup>
	2014	2.3 <sup>l</sup>	3.3 <sup>sl</sup>	4.0 <sup>sh</sup>	4.0 <sup>sh</sup>	5.1 <sup>h</sup>
Usefulness	2012	2.4 <sup>l</sup>	2.5 <sup>sl</sup>	4.1 <sup>sh</sup>	4.4 <sup>sh</sup>	5.5 <sup>vh</sup>
	2013	2.3 <sup>l</sup>	2.5 <sup>sl</sup>	3.7 <sup>sh</sup>	4.3 <sup>sh</sup>	5.5 <sup>vh</sup>
	2014	2.4 <sup>l</sup>	2.4 <sup>l</sup>	4.0 <sup>sh</sup>	3.8 <sup>sh</sup>	5.4 <sup>h</sup>
Success	2012	3.0 <sup>sl</sup>	4.6 <sup>h</sup>	4.0 <sup>sh</sup>	5.2 <sup>h</sup>	5.7 <sup>vh</sup>
	2013	3.4 <sup>sl</sup>	4.9 <sup>h</sup>	4.1 <sup>sh</sup>	5.2 <sup>h</sup>	5.6 <sup>vh</sup>
	2014	2.5 <sup>sl</sup>	4.6 <sup>h</sup>	3.9 <sup>sh</sup>	5.3 <sup>h</sup>	5.6 <sup>vh</sup>
Interest	2012	2.0 <sup>l</sup>	3.2 <sup>sl</sup>	3.6 <sup>sh</sup>	4.5 <sup>h</sup>	5.4 <sup>h</sup>
	2013	2.4 <sup>l</sup>	3.0 <sup>sl</sup>	4.0 <sup>sh</sup>	4.6 <sup>h</sup>	5.5 <sup>vh</sup>
	2014	2.1 <sup>l</sup>	2.6 <sup>sl</sup>	3.9 <sup>sh</sup>	4.4 <sup>sh</sup>	5.4 <sup>h</sup>
Caring	2012	2.8 <sup>sl</sup>	5.4 <sup>h</sup>	3.5 <sup>sh</sup>	5.4 <sup>h</sup>	5.6 <sup>vh</sup>
	2013	3.8 <sup>sh</sup>	5.6 <sup>vh</sup>	3.7 <sup>sh</sup>	5.5 <sup>vh</sup>	5.6 <sup>vh</sup>
	2014	3.6 <sup>sh</sup>	5.4 <sup>h</sup>	3.9 <sup>sh</sup>	5.5 <sup>vh</sup>	5.6 <sup>vh</sup>
Cluster N (%)	2012	29 (8.95%)	37 (11.42%)	56 (17.28%)	92 (28.40%)	107 (33.02%)
	2013	45 (14.61%)	40 (12.99%)	54 (17.53%)	92 (29.87%)	77 (25.00%)
	2014	24 (8.45%)	33 (11.62%)	48 (16.90%)	75 (26.41%)	104 (36.62%)

Note. All variables were significantly different between all clusters,  $p < .001$ . 2012  $n = 321$ ; 2013  $n = 308$ ; 2014  $n = 284$ ;  $N = 913$ .  
<sup>vl</sup> Very low = 1.0 to 1.4. <sup>l</sup> Low = 1.5 to 2.4. <sup>sl</sup> Somewhat low = 2.5 to 3.4. <sup>sh</sup> Somewhat high = 3.5 to 4.4. <sup>h</sup> High = 4.5 to 5.4. <sup>vh</sup> Very high = 5.5 to 6.0.

though parsimonious, did not adequately describe the data, providing only “high,” “middle,” and “low” clusters in which the nuances of motivation were lost.

**Initial cluster typology.** Cluster solutions should also be interpretable with names and classifications informed by theory (Bacher, 2002). To define the cluster profiles, we first organized

the cluster centers into six categories that described the students’ reported perceptions, per the 6-point scale: very low (1.0 to 1.4), low (1.5 to 2.4), somewhat low (2.5 to 3.4), somewhat high (3.5 to 4.4), high (4.5 to 5.4), and very high (5.5 to 6.0). We selected terminology that describes an amount or quantity of science class perceptions, similar to Daniels et al.’s (2008), Vansteenkiste et

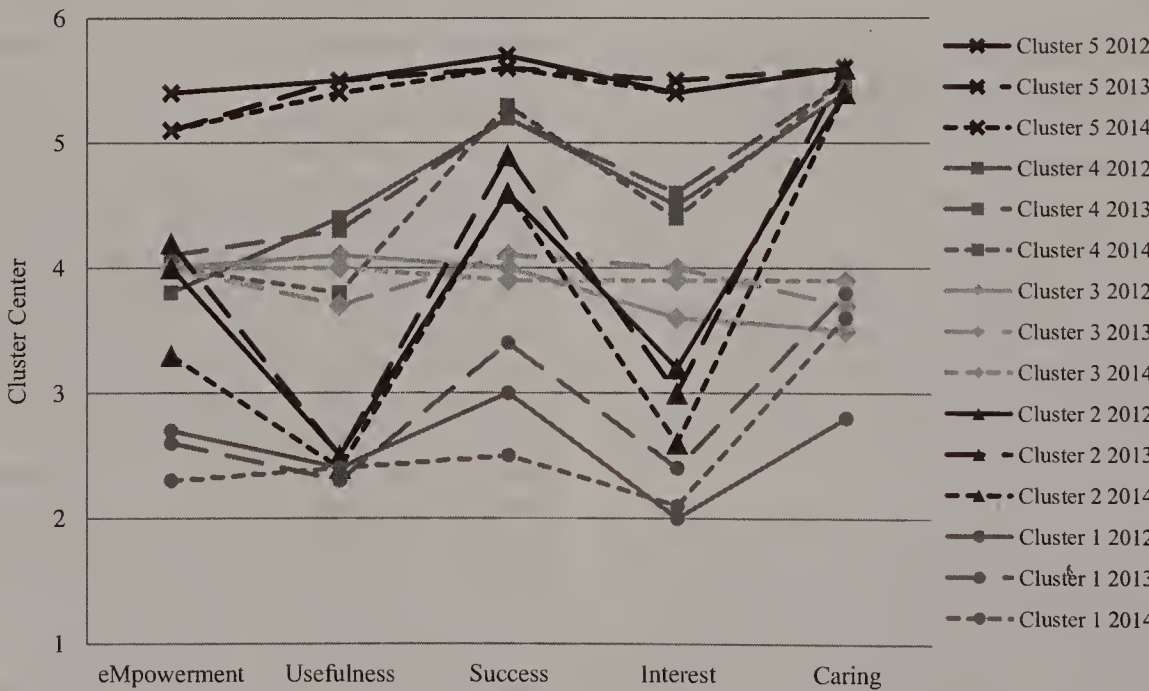


Figure 2. Cluster centers for each year (2012 to 2014). This figure shows how the different clusters are stable across the years. The five clusters are differentiated by different shades and marker styles: black with “X” marker = Cluster 5; dark gray with square marker = Cluster 4; light gray with diamond marker = Cluster 3; black with triangle marker = Cluster 2; dark gray with circle marker = Cluster 1. Years are demarcated with different lines: solid = 2012; large dashes = 2013; small dashes = 2014.



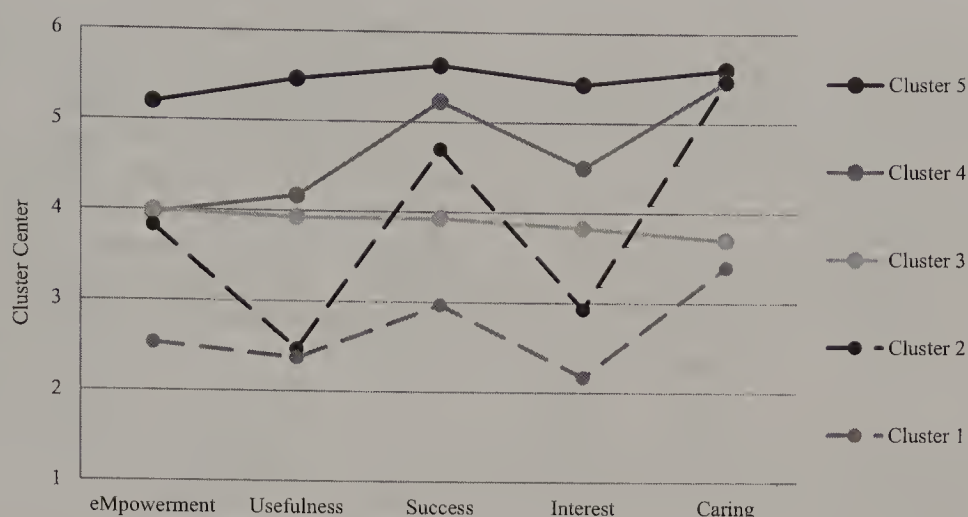


Figure 3. Collapsed cluster centers based on year. This figure simplifies the visual of cluster centers. Each line represents the mean of the three years (2012, 2013, 2014) for the five MUSIC model components, per cluster. The five clusters are differentiated by different shades and line styles: black solid line = Cluster 5; dark gray solid line = Cluster 4; light gray solid line = Cluster 3; black dashed line = Cluster 2; dark gray dashed line = Cluster 1.

al.'s (2009), Wormington et al.'s (2012), and Hayenga and Corpus's (2010) conception of low or high "quantity" motivation. The items in the MUSIC Inventory measure the *quantity* of students' perceptions about science class rather than the *quality* of those perceptions.

Using the very low to very high categories to explain each variable's cluster center within the overall cluster membership, our initial characterization of the five clusters was as follows: (a) low motivation; (b) low usefulness and interest, moderate empowerment, and high success and caring; (c) somewhat high motivation; (d) somewhat high empowerment, usefulness, and interest, and high success and caring; and (e) high motivation. Findings for these analyses are displayed in Table 5. The average percentage of students per cluster was similar across years, with the majority of students (55% to 63%) assigned to Clusters 4 and 5. One-way ANOVAs indicated that all five clusters were significantly different on the clustered variables ( $p < .001$ ), which is expected because the purpose of cluster analysis is to maximize within-cluster homogeneity and between-cluster heterogeneity. Intercorrelations were also low ( $-.006$  to  $.110$ ), supporting this conclusion.

**Discriminant analysis.** To further distinguish each motivation profile, we computed a discriminant factor analysis (Burns & Burns, 2008) with the 2014 dataset ( $n = 284$ ). We used the 2014 dataset as an exemplar because the motivation profiles were relatively similar across years. Four functions emerged, which was expected because the maximum number of functions possible is

the number of clusters minus one (Burns & Burns, 2008). Table 6 includes structure coefficients for the four functions. Function 1 ( $D_1$ ) is the dominant function, as it explained 83% of the between-groups variance and, together,  $D_1$  and  $D_2$  explained 98.5%.  $D_3$  and  $D_4$  were responsible for a negligible amount of explained variance (1.6% combined) and were not key factors in cluster membership for any of the five profiles. Accordingly, and with our descriptive intention in mind, we omitted  $D_3$  and  $D_4$  from these results to maintain a parsimonious model.

We interpreted  $D_1$  and  $D_2$  using the discriminant loadings, which are Pearson correlations between the functions and MUSIC model variables, and indicate which variables were most important or influential within each function (Burns & Burns, 2008).  $D_1$  is associated with a high level of interest, success, and usefulness (in this order). Empowerment and caring were considered less critical factors.  $D_2$  is primarily associated with a high level of perceived caring (the most important predictor), as well as with high perceived success and low usefulness, which were similarly weighted. Neither empowerment nor interest was significant in this model.

Table 7 shows the discriminant functions at each cluster centroid. The discriminant function coefficient at Cluster 1 indicates very low interest, success, and usefulness ( $-5.78$ ), indicating that students' very low perceptions of these MUSIC components were prominent influences of membership in the low motivation profile. Empowerment was not a meaningful factor in the Cluster 1 membership and associations with all other functions were low. Cluster 2 membership suggests low interest, success, and usefulness

Table 6  
Discriminant Function Analysis, 2014

Function	Eigenvalue	% of variance	Canonical correlation	$\chi^2$	df	Interpretation
$D_1$	7.165	83.1	.937	853.7	20	High interest, success, and usefulness
$D_2$	1.324	15.4	.755	269.9	12	High caring and success, low usefulness
$D_3$	0.082	1.0	.275	35.5	6	—
$D_4$	0.050	0.6	.219	13.6	2	—

Table 7  
Unstandardized Canonical Discriminant Functions at Cluster Centroid, 2014

Cluster	Function			
	D <sub>1</sub> High interest, success, and usefulness	D <sub>2</sub> High caring and success, and low usefulness	D <sub>3</sub>	D <sub>4</sub>
1	<b>-5.783</b>	-0.881	0.029	0.519
2	<b>-2.817</b>	<b>2.087</b>	0.436	-0.195
3	<b>-1.614</b>	<b>-1.733</b>	-0.033	-0.331
4	0.268	0.855	-0.422	-0.004
5	<b>2.780</b>	-0.276	0.175	0.098

Note. Bold font indicates the most important factors to cluster membership.

(-2.82) and, at the same time, high caring and success, and low usefulness (2.09). Cluster 2 is consistent with a profile with high caring, moderate success, and very low interest and usefulness. We indicated a moderate level of success because the success variable was contradictory between functions, with low success in D<sub>1</sub> and high success in D<sub>2</sub>, and the canonical discriminant function coefficients were similarly weighted. Empowerment was not a meaningful factor in the Cluster 2 membership and influences of other functions were low. Cluster 3 was moderately low on the high interest and success, and low usefulness factor (-1.61), and moderately low on the high caring and success, and low usefulness factor (-1.73). Thus, Cluster 3 suggests that these students held fairly moderate to somewhat high perceptions of *all* variables, and that empowerment was not an influential factor in the Cluster 3 membership. Cluster 4 indicates that no single variable or factor was especially influential in cluster membership in that no function

was particularly significant; rather, the similar correlation coefficients suggest a combination of several influential variables. Finally, students in Cluster 5 indicated high interest, success, and usefulness (2.78), which was more important to their cluster membership. These findings are the inverse of Cluster 1, the “low motivation” profile, in which extremely low perceptions of interest, success, and usefulness held weight.

**Final cluster typology.** Combining results of the discriminant analysis with our earlier categorization based on the cluster centers of each MUSIC model variable, we developed the following labels to describe each cluster: (a) *low motivation*; (b) *low usefulness and interest, but high success and caring*; (c) *somewhat high motivation*; (d) *somewhat high motivation, and high success and caring*; and (e) *high motivation*.

**Stage II: Stability tests.** To test the stability of the clusters when organized into different subsets, we followed the same two-step clustering procedure for separate grade levels (fifth *n* = 324, sixth *n* = 263, seventh *n* = 326) rather than years. These stability tests were intended to reduce the teacher effect and effects of unknown and contextual variables. We found that the cluster solution and cluster centers remained stable. The five-cluster solution best fit each grade level and the cluster centers aligned with the 2012, 2013, and 2014 clusters. See Table 8 for the cluster centers by grade level and Figures 4 and 5 for visual representations of the five clusters and their stability at each grade level.

**Gender and grade level associations.** Given previous research citing gender and age effects on science motivation (Bong et al., 2015; Maltese & Tai, 2010; Meece et al., 2006), we investigated gender and grade level differences among motivation profiles. Pearson chi-square tests revealed significant differences between genders,  $\chi^2(4, N = 913) = 13.45, p = .001$  and grade levels,  $\chi^2(8, N = 913) = 48.01, p < .001$  (see Figures 6 and 7). A higher

Table 8  
Five-Cluster Solution: Comparisons Among Grade Levels

MUSIC component	Year	Clusters				
		1	2	3	4	5
Empowerment	5th	2.2 <sup>l</sup>	3.5 <sup>sh</sup>	4.0 <sup>sh</sup>	4.2 <sup>sh</sup>	5.4 <sup>h</sup>
	6th	2.8 <sup>sl</sup>	3.8 <sup>sh</sup>	4.1 <sup>sh</sup>	4.0 <sup>sh</sup>	5.1 <sup>h</sup>
	7th	2.8 <sup>sl</sup>	2.9 <sup>sl</sup>	4.0 <sup>sh</sup>	3.9 <sup>sh</sup>	5.1 <sup>h</sup>
Usefulness	5th	2.5 <sup>sl</sup>	2.3 <sup>l</sup>	3.9 <sup>sh</sup>	4.5 <sup>h</sup>	5.6 <sup>vh</sup>
	6th	2.3 <sup>l</sup>	2.5 <sup>sl</sup>	3.9 <sup>sh</sup>	4.2 <sup>sh</sup>	5.4 <sup>h</sup>
	7th	2.9 <sup>sl</sup>	1.8 <sup>l</sup>	4.0 <sup>sh</sup>	3.8 <sup>sh</sup>	5.3 <sup>h</sup>
Success	5th	2.5 <sup>sl</sup>	4.6 <sup>h</sup>	4.2 <sup>sh</sup>	5.3 <sup>h</sup>	5.7 <sup>vh</sup>
	6th	3.0 <sup>sl</sup>	4.7 <sup>h</sup>	3.5 <sup>sh</sup>	5.0 <sup>h</sup>	5.7 <sup>vh</sup>
	7th	3.0 <sup>sl</sup>	4.2 <sup>sh</sup>	4.3 <sup>sh</sup>	5.2 <sup>h</sup>	5.6 <sup>vh</sup>
Interest	5th	2.3 <sup>l</sup>	3.0 <sup>sl</sup>	3.9 <sup>sh</sup>	4.8 <sup>h</sup>	5.5 <sup>vh</sup>
	6th	2.1 <sup>l</sup>	3.1 <sup>sl</sup>	4.0 <sup>sh</sup>	4.3 <sup>sh</sup>	5.5 <sup>vh</sup>
	7th	2.3 <sup>l</sup>	2.2 <sup>l</sup>	3.9 <sup>sh</sup>	4.2 <sup>sh</sup>	5.3 <sup>h</sup>
Caring	5th	3.0 <sup>sl</sup>	5.6 <sup>vh</sup>	3.9 <sup>sh</sup>	5.6 <sup>vh</sup>	5.8 <sup>vh</sup>
	6th	3.5 <sup>sh</sup>	5.3 <sup>h</sup>	3.6 <sup>sh</sup>	5.3 <sup>h</sup>	5.5 <sup>vh</sup>
	7th	2.7 <sup>sl</sup>	5.2 <sup>h</sup>	3.6 <sup>sh</sup>	5.5 <sup>vh</sup>	5.5 <sup>vh</sup>
Cluster <i>N</i> (%)	5th	15 (4.62%)	42 (12.92%)	50 (15.38%)	113 (34.56%)	105 (32.31%)
	6th	30 (11.36%)	40 (15.15%)	32 (12.12%)	75 (28.41%)	87 (32.95%)
	7th	42 (12.84%)	34 (10.40%)	66 (20.18%)	94 (28.75%)	91 (27.83%)

Note. All variables were significantly different between all clusters, *p* < .001. Fifth-grade *n* = 324; sixth-grade *n* = 263; seventh-grade *n* = 326; *N* = 913.

<sup>sl</sup> Very low = 1.0 to 1.4. <sup>l</sup> Low = 1.5 to 2.4. <sup>sl</sup> Somewhat low = 2.5 to 3.4. <sup>sh</sup> Somewhat high = 3.5 to 4.4. <sup>h</sup> High = 4.5 to 5.4. <sup>vh</sup> Very high = 5.5 to 6.0.



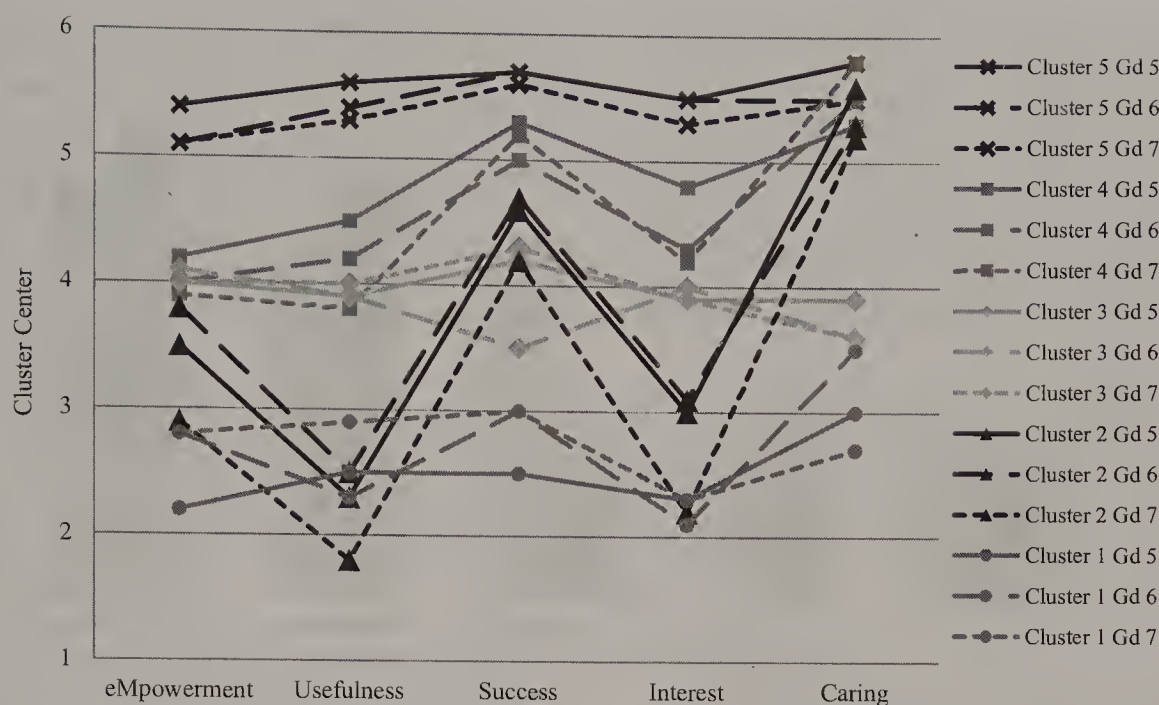


Figure 4. Cluster centers for each grade level (fifth, sixth, seventh). This figure shows how the different clusters are stable across the grade levels. The five clusters are differentiated by different shades and marker styles: black with "X" marker = Cluster 5; dark gray with square marker = Cluster 4; light gray with diamond marker = Cluster 3; black with triangle marker = Cluster 2; dark gray with circle marker = Cluster 1. Grade levels are demarcated with different lines: solid = fifth-grade; large dashes = sixth-grade; small dashes = seventh-grade.

proportion of female students were assigned to Clusters 4 and 5, and a lower proportion were in Cluster 3. Inversely, males were overrepresented in Cluster 3 and underrepresented in Cluster 5. Fifth-grade students were overrepresented in Cluster 5 and underrepresented in Clusters 1 and 3. Inversely, seventh-grade students were overrepresented in Clusters 1 and 3, and underrepresented in Cluster 5. Sixth-grade students were underrepresented in Cluster 5.

### Students' Cluster Membership Across Years

To test whether or not students remained in the same clusters every year, we selected the 167 students who completed the

questionnaire at more than one time point (i.e., at 2012 and 2013, and/or 2012 and 2014, and/or 2013 and 2014), and we ran a series of Cohen's  $\kappa$  tests to compare their cluster memberships between years. We found that students' cluster memberships varied across years: between 2012 and 2013,  $\kappa = .191$  ( $n = 149$ ); between 2012 and 2014,  $\kappa = .135$  ( $n = 65$ ); and between 2013 and 2014,  $\kappa = .290$  ( $n = 47$ ). We were able to test a total of 261 comparisons between two separate years (2012/2013, 2012/2014, 2013/2014) because some of the 167 students were assessed all three years. Examined another way, of the students for which we had data spanning more than one year, 37.1% did not move to a new cluster,

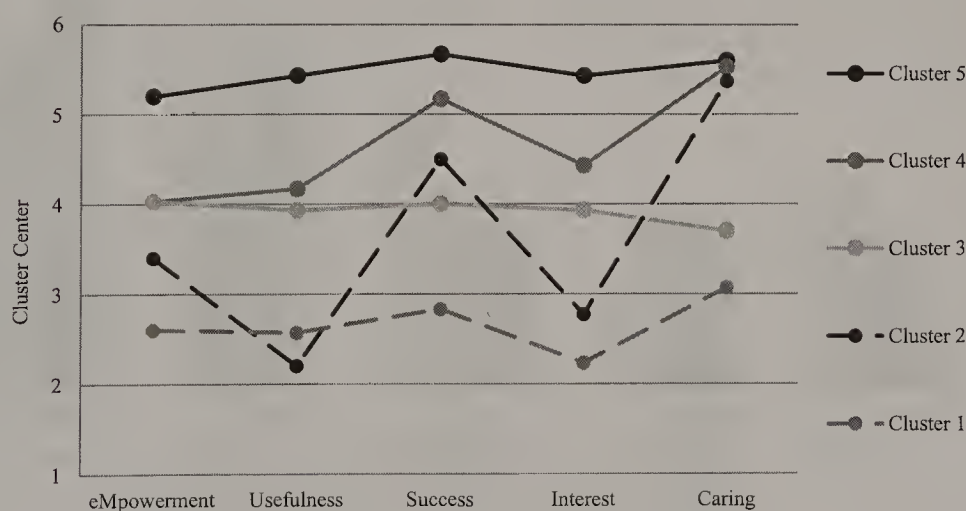


Figure 5. Collapsed cluster centers based on year. This figure simplifies the visual of cluster centers. Each line represents the mean of the three years (2012, 2013, and 2014) for the five MUSIC model components, per cluster. The five clusters are differentiated by different shades and line styles: black solid line = Cluster 5; dark gray solid line = Cluster 4; light gray solid line = Cluster 3; black dashed line = Cluster 2; dark gray dashed line = Cluster 1.

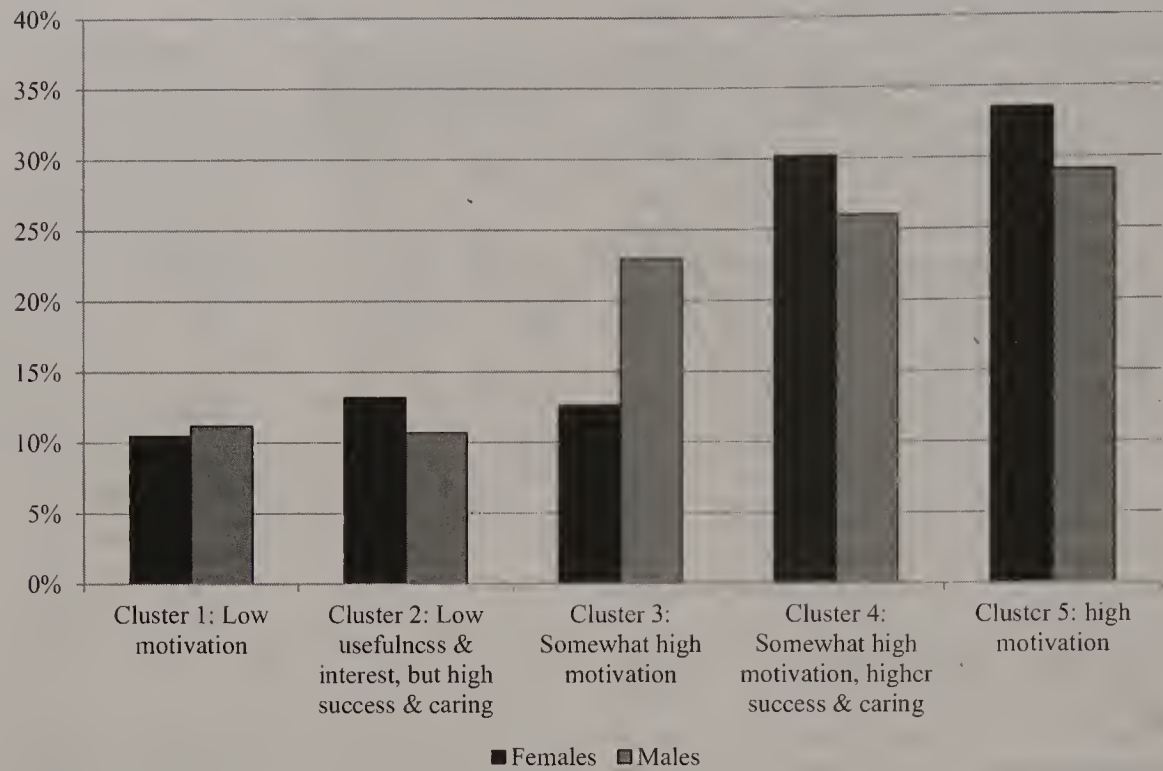


Figure 6. Gender distribution in clusters. This figure includes the percentage of each gender out of the full sample that was categorized into each cluster. Female  $n = 485$ ; male  $n = 428$ .

40.9% moved to a lower cluster number (e.g., Cluster 3 to Cluster 1), and only 21.8% moved to a higher cluster number (e.g., Cluster 3 to 5) over time. These findings suggest that cluster membership may be somewhat dependent on the context of each specific science class, science teacher, or some other variable or combination of variables.

Predictive Validity

To provide more evidence for the predictive validity of our cluster solution (Bacher, 2002), we completed several follow-up tests on the 2014 data with variables that have been shown to correlate theoretically and empirically: (a) identification for sci-

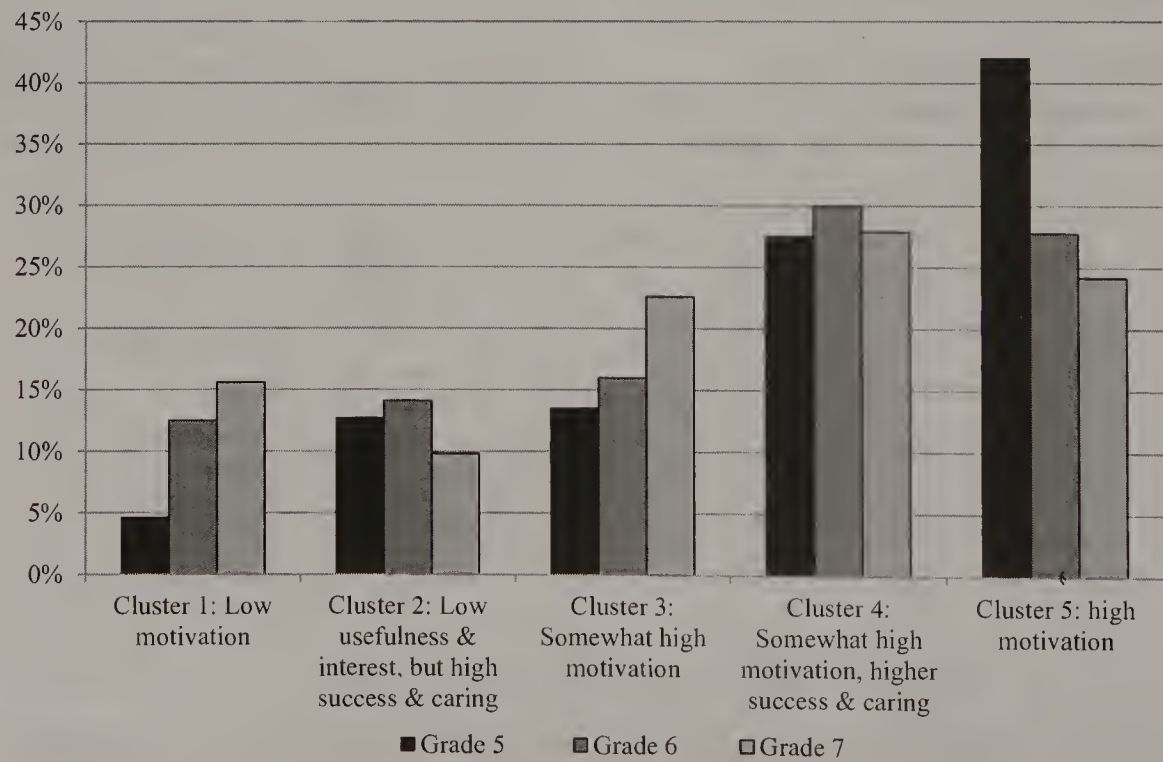


Figure 7. Grade level distribution among clusters. Because the number of students in each grade level was uneven, this figure shows the percentage of students in the grade level that was grouped into each cluster. Grade 5  $n = 324$ ; Grade 6  $n = 263$ ; Grade 7  $n = 326$ .



ence, (b) science career goals, (c) science course intentions, and (d) science class effort. We selected the 2014 data ( $n = 284$ ) to act as exemplar, as the motivation profiles and ANOVA results were similar across years. By selecting a single year, we sought to reduce complexity and present the information in a coherent manner. One-way ANOVAs revealed significant differences among motivation profiles in science class and each outcome variable (see Table 9). We computed Tukey's HSD and Games-Howell post hoc tests for all variables, which we describe next (Table 10 and Figure 8).

**Science identification.** Post hoc tests revealed significantly higher reported science identification in Cluster 5 than in all other clusters. Cluster 4 included significantly higher reported science identification than Clusters 3, 2, and 1. Clusters 3 and 2 included statistically similar reported science identification, which was significantly higher than in Cluster 1. Cluster 1 had the lowest reported science identification.

**Science career goals.** Post hoc tests indicated significantly higher reported science career goals in Cluster 5 than in all other clusters. Clusters 4, 3, and 2 included statistically similar reported career goals, as did Clusters 2 and 1, albeit lower. Cluster 1 had the lowest reported science career goals, which were significantly lower than all other clusters except Cluster 2.

**Science course intentions.** Post hoc tests revealed that Cluster 5 included significantly higher reported intentions to take science courses in the future than all other clusters. Clusters 4 and 3 included statistically similar reported course-related intentions, as did Clusters 1 and 2; however, Clusters 1 and 2 were significantly lower than all other clusters.

**Science class effort.** Post hoc tests revealed that Cluster 5 included higher reported science class effort than all other clusters, and Cluster 4 was higher than Clusters 3, 2, and 1. Clusters 3 and 2 included statistically similar reported effort in science class, and Cluster 1 had significantly lower reported effort than all other clusters.

## Discussion

Our objective was to use a person-centered approach to categorize students into multidimensional motivation profiles in science class based on five well-known motivation constructs that have been shown to relate to students' science identification and inten-

tions to persist in science. Our use of person-centered analyses allowed us to identify complex patterns and offer a more inclusive view of students' motivation than more commonly used linear research methods (Meece & Holt, 1993). To our knowledge, researchers have not yet used cluster analysis to examine the motivation of pre-high school science students in this manner. Our study demonstrates that patterns of science class perceptions form five stable clusters, which are theoretically meaningful and depict the students' multidimensional class-related motivation profiles.

We describe the five profiles in more detail next, including their associations with other measures (as shown in Figure 8), to provide evidence of predictive validity. We classified each cluster according to the quantity of motivation reported (i.e., cluster centers and descriptive statistics) and the factors that were most important to cluster membership, per the discriminant analysis.

## Profile Descriptions

**Cluster 1: Low motivation profile.** Students in Cluster 1 reported a lower quantity of motivation for science class than students in the other clusters. The motivation profile for these students is primarily characterized by very low perceived interest, success, and usefulness. Although less influential in determining cluster membership, they also perceived low empowerment in science class and felt that their teachers were only moderately caring.

Of the five clusters, the students in Cluster 1 reported that they identified the least with science and applied the least amount of effort to their science classes. Likewise, they reported little intention to persist in science, either by taking future science courses or considering the pursuit of a science-related career. These findings are consistent with previous research positing that students with low motivation tend to put forth little effort, have low expectancy for success, and lack value for the subject (Legault, Green-Demers, & Pelletier, 2006).

**Cluster 2: Low usefulness and interest, but high success and caring profile.** Students in Cluster 2 reported that they held low usefulness and interest and had little control over their learning in science class; however, they expected to be successful and felt cared for in the classroom. This profile is primarily characterized

Table 9  
*One-Way ANOVA Results for Outcome Variables*

	<i>M (SD)</i>		<i>SS</i>	<i>df</i>	<i>MS</i>	<i>f</i>
Science identification	4.52 (1.11)	Between groups	156.764	4	39.191	57.036*
		Within groups	191.707	279	0.687	
		Total	348.471	283		
Science career goals	3.39 (1.57)	Between groups	176.891	4	44.223	23.871*
		Within groups	516.865	279	1.853	
		Total	693.756	283		
Science course intentions	3.30 (1.70)	Between groups	225.828	4	56.457	26.728*
		Within groups	589.327	279	2.112	
		Total	815.155	283		
Science class effort	4.74 (1.16)	Between groups	151.744	4	37.936	60.878*
		Within groups	173.857	279	0.623	
		Total	325.601	283		

\*  $p < .001$ .

Table 10  
Means and Standards Deviations per Cluster for Each Outcome Variable

Variable	Cluster 1 <i>M (SD)</i>	Cluster 2 <i>M (SD)</i>	Cluster 3 <i>M (SD)</i>	Cluster 4 <i>M (SD)</i>	Cluster 5 <i>M (SD)</i>
Science identification	2.88 (0.90)	3.94 (1.08)	3.93 (0.84)	4.63 (0.75)	5.30 (0.68)
Science career goals	2.06 (1.46)	2.53 (1.32)	3.19 (1.30)	2.99 (1.20)	4.35 (1.48)
Science course intentions	1.88 (1.26)	2.06 (1.44)	3.00 (1.43)	3.01 (1.39)	4.36 (1.55)
Science class effort	3.12 (1.17)	4.11 (0.88)	4.18 (0.89)	4.82 (0.80)	5.50 (0.54)

by high perceived caring, moderate success expectancies, and very low situational interest and perceived usefulness.

Students in Cluster 2 reported that they moderately identified with science and put forth a moderate amount of effort in science class, similar to Cluster 3. However, their desire to take more science classes and/or to pursue science-related careers was low. When students do not intend to persist, they tend to report lower values like usefulness and interest (Meece, Wigfield, & Eccles, 1990; Schunk & Pajares, 2005), and less autonomy (Deci & Ryan, 2000; Schunk, 1995), such as these students reported for their science classes. Overall, these students may not have perceived science to be useful or experienced much enjoyment and situational interest in science class, and felt little empowerment with respect to their learning, but they asserted some effort in class, expected to do fairly well, and believed that their teachers cared about their academic and personal well-being. Even though these students did not have many intentions to persist in science, it is possible that their moderate effort in class could have been motivated by other factors, such as their expectations for success (Cox & Whaley, 2004; Gendolla, Wright, & Richter, 2012; Greene, DeBacker, & Krows, 1999; Pajares, 1996), high perceptions of teacher caring (Wentzel, 1997), and/or more external factors we did not measure, such as attaining a specific grade.

**Cluster 3: Somewhat high motivation profile.** Students in Cluster 3 reported that they perceived only somewhat high levels of empowerment, usefulness, success, interest, and caring in their science classes. We propose that a lack of highly positive caring and success beliefs differentiates this profile from Clusters 2 and 4.

Students in Cluster 3 reported moderate science identification and science class effort, similar to Cluster 2; however, unlike Cluster 2, they reported only somewhat low intentions to persist in science. Overall, these data suggest that somewhat high motivation for science class in this case seem to be associated with similarly moderate to moderately low science identification, intentions to persist in science, and effort put forth in class.

**Cluster 4: Somewhat high motivation, and high success and caring profile.** Cluster 4 reported a somewhat high amount of empowerment, moderate to high perceptions of usefulness and interest, and that they expected to be very successful and perceived a high level of caring in science class. However, Cluster 4 expressed somewhat low science career goals and science course intentions. Regardless of their low intentions to persist in the field, they valued science as an important part of who they were and, likewise, put forth a lot of effort in science class. Overall, these students held only moderate perceptions of usefulness and interest, and it seems, corresponding to these beliefs, they had few plans to

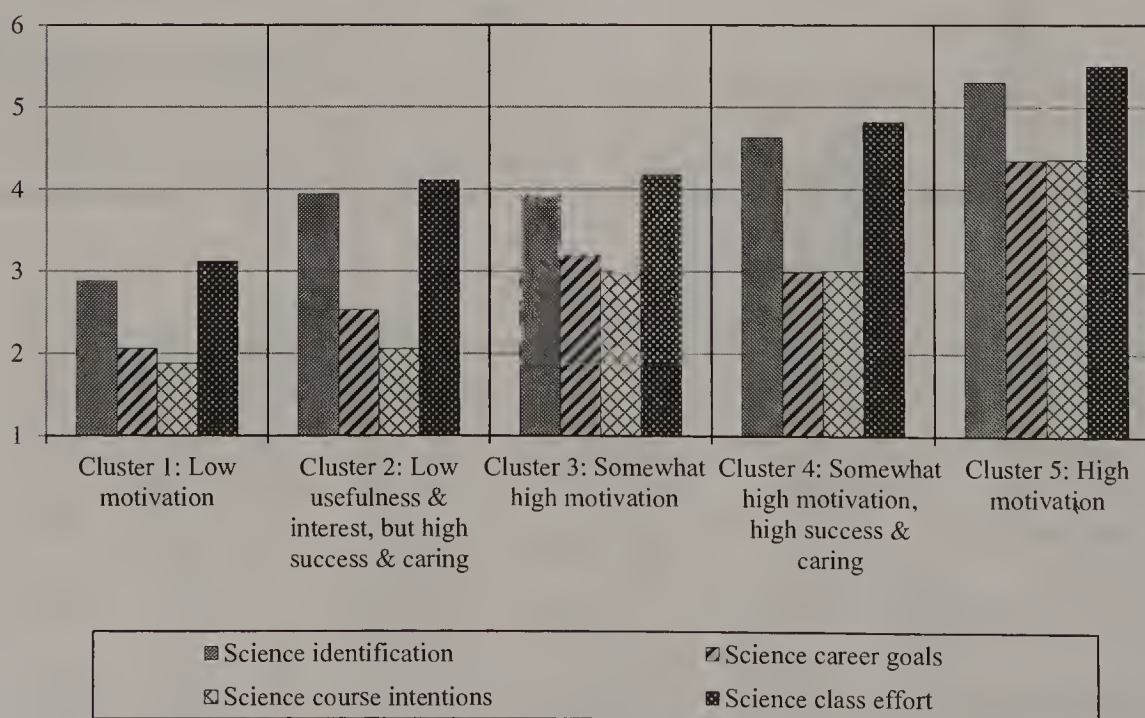


Figure 8. Cluster comparisons across correlated variables (2014 data set). Each bar represents one variable and each section represents one cluster.



pursue science courses and careers; nevertheless, they placed importance on doing well in science, tried hard, expected to be successful, and believed that their teachers cared.

We more closely examined dissimilarities between Clusters 2 and 4, and Clusters 3 and 4, respectively. In general, it appears that the key distinctions between Clusters 2 and 4 were lower perceived usefulness and interest in science class in Cluster 2, which were negatively associated with reported effort and identification. This is evidenced in several ways. First, mean values for each science class perception indicates that the profiles are similar except that in Cluster 4, perceived usefulness and interest in science class were much higher, and success expectancies appeared only slightly higher. Also, perceived empowerment in science class and science career goals were similar between profiles; however, those in Cluster 2 reportedly extended significantly less effort in science class, and their science identification and intentions to take future science courses were notably lower.

Cluster 4 is primarily different than Cluster 3 in that perceived caring and success expectancies in science class were higher in Cluster 4. Differences in usefulness and interest appear more minor, although Cluster 4 was generally more interested during science class. Cluster 4 reported similar intentions to persist as Cluster 3; however, they reportedly put forth much more effort in science class and indicated significantly higher science identification. Thus, these data may imply a positive relationship between high caring and success expectancies, and student's effort in science class and domain identification.

**Cluster 5: High motivation profile.** Students in Cluster 5 reported higher quantity motivation for science class than students in the other clusters. Discriminant analysis revealed that their motivation for science class was primarily characterized by high situational interest, success expectancies, and perceived usefulness. Although less influential in determining cluster membership, they also generally perceived that they had control over their learning in science class and that their teachers were highly caring. Students in this cluster reported significantly greater perceptions on all four outcome variables: they believed that science was an important part of their identities, they put forth a lot of effort in science class, they reported that they wanted to take science classes in the future, and they could see themselves in science-related careers. These findings are consistent with previous research, which suggests that students with high motivation are often more engaged in class, put forth more effort, hold positive motivation-related beliefs, and intend to persist in the future (Deci & Ryan, 2000; Hidi & Renninger, 2006; Wigfield & Eccles, 2000). It is important to note that Clusters 1 through 4 reported low to somewhat low intentions to persist; of all clusters, only Cluster 5 reported somewhat high intentions to persist.

## Trends Across Profiles

We identified several important trends that can help to further unravel students' motivations. First, we found that situational interest and usefulness tended to be lower than perceptions of success and caring in science class for most profiles. Only students in the high motivation profile reported high situational interest and usefulness for science class. These patterns may indicate that, although students in Clusters 2 and 4 believe that they can be successful and that their teachers are caring, they may not have any

particular desire to engage in science activities, unlike those in Cluster 5 (e.g., Eccles, Wigfield, & Schiefele, 1998). These results may also imply that situational interest and usefulness are not as critical as perceptions of success in terms of the effort students put forth in class and their identification with science, as multiple clusters were associated with somewhat high to high reports of effort in science class and/or science identification. This finding aligns with literature that associates success expectancies and effort (Cox & Whaley, 2004; Gendolla et al., 2012; Greene et al., 1999; Pajares, 1996), and success expectancies and identification (Osborne & Jones, 2011). Only the cluster with high perceived situational interest and usefulness in science class (Cluster 5) also reported somewhat high intentions to persist, which is consistent with previous findings that perceptions of usefulness are positively associated with persistence (Jones et al., 2010; Jones, Tendhar, & Paretti, 2016; Meece et al., 1990). Furthermore, lack of situational interest and perceived usefulness has been related to amotivation, attrition, and other negative outcomes (Ryan & Deci, 2000; Legault et al., 2006; Renninger, Hidi, & Krapp, 1992), which appears consistent with the low to somewhat low intentions to persist in science reported by all clusters except Cluster 5. As in Cluster 5, there is evidence indicating that a combination of high situational interest, success, and usefulness is positively associated with persistence, effort, course selection, and choice of college major (Aschbacher, Ing, & Tsai, 2014; Eccles et al., 1983)—outcomes especially salient in the present study. We propose that students in Cluster 5 are more likely to persist in science and perform well in the future, and their reported intentions seem to suggest the same. We do not present evidence to explain whether perceived situational interest and usefulness in science class were lower in Clusters 1 through 4 or whether perceived success and caring were simply higher (i.e., students perceived science class as easier and believed that their teachers were caring). Additional research is needed to better understand the implications of these findings.

Second, we also found noteworthy trends concerning empowerment. A high level of perceived empowerment in science class was noted only in Cluster 5. This finding is consistent with self-determination theory (Deci & Ryan, 2000), which states that motivation associated with more positive outcomes is considered more autonomous (i.e., internalized motivation and intrinsic motivation; Deci & Ryan, 2012; Ryan, 1995) and higher quality (Vansteenkiste et al., 2006). Prior investigations indicated that students who are given some autonomy and hold high situational interest and usefulness for the task—as in Cluster 5—are more likely to function positively (e.g., increased engagement; Assor, Kaplan, & Roth, 2002), and that provision of fewer choices has predicted decreased interest and usefulness perceptions (e.g., Midgley & Feldlaufer, 1987). Furthermore, the measure we used for situational interest in science class is analogous to some measures of intrinsic motivation (e.g., Reeve, 1989), implying that students in Cluster 5 both perceived more empowerment in science class and were more autonomously motivated than in all other profiles. However, empowerment did not present as an influential variable in discriminant analysis, indicating that it may not be as important as the other variables in categorizing students into a particular profile. In other words, perceived autonomy was not the organizing fact of clusters in this study, which seems to conflict with some arguments that meeting the need for autonomy is particularly important to high quality motivation (Deci & Ryan,



1985, 2012). Still, we measured *quantity* of motivation with the MUSIC Inventory (i.e., the amount of control and freedom the students felt they experienced in science class) rather than quality. The particular significance of autonomy in self-determination theory centers on the concept of higher *quality* motivation (Vansteenkiste et al., 2006), which is considered more autonomous or, in other words, volitional, valued, and endorsed by the sense of self (Deci & Ryan, 2012; Vansteenkiste et al., 2006). Students identified as part of Cluster 5, which would be the profile most indicative of a high *quality* motivation profile, reported the most perceived empowerment in class. These students not only reported a high amount of perceived control in their science classes—unlike all other profiles—but also reported the highest perceived situational interest, expectancies for success, and usefulness in science class, as well as the highest science identification, effort in science class, and intentions to persist in the field. These findings require further investigation, as we did not measure the direction of the relationships in the present study nor do we attempt to assert causal relationships.

Third, although post hoc assessment of differences among the MUSIC components across profiles is considered an inappropriate test given the nature of cluster analysis, a distinction in the caring perceptions across profiles is visibly discernible. There appear to be two fairly consistent “groupings,” per se, for caring: one that indicates high to very high perceived caring in science class (found in Clusters 2, 4, and 5) and a second indicating somewhat low to somewhat high perceived caring in science class (Clusters 1 and 3). Of the five Clusters, each fell into one of these two groups, indicating that students either tended to think their science teachers very caring or somewhat caring, with few in between. It is beyond the scope of the present study to further evaluate the cause of this trend; however, it appears to suggest that students in this population may be prone to somewhat dichotomous perceptions of teacher caring.

## Gender Differences

Disproportionately more female students were assigned to Clusters 4 and 5 than to Clusters 1, 2, and 3. This finding is consistent with the results of previous cluster analyses that indicated a higher proportion of female students in similarly high quantity motivation profiles (Ratelle et al., 2007; Wormington et al., 2012) and some investigations of secondary level students more generally (Fischer, Schult, & Hell, 2013). Our findings (see Figure 7) indicate that female students were generally more motivated in science classes than their male counterparts, which also contradicts some research we cited previously that suggests male students are often more motivated in science courses (Bong et al., 2015; Eccles, 2007; Maltese & Harsh, 2015; Meece et al., 2006).

## Grade Level Differences

Fifth-grade students were overrepresented in Cluster 5 and underrepresented in Clusters 1 and 3 (see Figure 6). This finding generally suggests that fifth-grade students were more highly motivated than the older students. In a direct inverse of this finding, seventh-grade students were underrepresented in Cluster 5, and overrepresented in Clusters 1 and 3, suggesting that the oldest students were less motivated for science class than their

younger peers. Similarly, there were fewer than expected sixth-grade students in the high motivation cluster. Together, these findings are consistent with previous studies indicating that motivation in science often declines with age (Eccles et al., 1993; Jacobs et al., 2002).

## Context-Dependent Motivation

Of those students who completed the survey at multiple time points, few retained the same profile during two or more years. It is important to note that the questionnaire items specifically targeted perceptions of the students' current science classes. Thus, this finding supports the notion that motivation-related perceptions often depend on the specific context of each class, which previous research suggests can be influenced by teachers and the educational environment (e.g., Dotterer & Lowe, 2011; Neiswandt & Shanahan, 2008; Steinmayr & Spinath, 2008; Urdan & Schoenfelder, 2006; Wang & Eccles, 2013). If so, these findings may underscore the importance of teachers' behaviors and instructional design in affecting students' motivation in science classes.

We posit that the changes in students' motivation profiles are not attributable to science curriculum differences per grade level because the curriculum at the schools in this study encompassed two to three science disciplines per grade level, and physical science—sometimes considered the most rigorous and difficult—was taught in all three grade levels. In all, this is significant information, as it may encourage educators to view students' motivation as phenomena that can change from year-to-year and possibly class-to-class, rather than remain fixed and unchangeable. However, more research is needed to support this notion, as other factors may have contributed to the transient cluster membership (e.g., home life, age/grade level, environmental influences on testing days).

We further suggest that the changes in students' profiles may be related, in part, to the typical waning of students' motivation over time, given that 40.9% moved to a lower cluster number, 37.1% did not move to a new cluster, and only 21.8% moved to a higher cluster number. If students' science motivation typically decreases over time (Osborne et al., 2003; Simpson & Oliver, 1990), the nature of these profiles would be such that we would expect students to move from higher cluster numbers to lower cluster numbers between years. We could expect this because, compared with the higher cluster numbers, students in the lower cluster numbers reported lower or similar science class perceptions (with the exception that the Cluster 2 scores for perceived success and caring are higher than the Cluster 3 scores), as shown in Figure 3. These yearly changes highlight the importance of teachers and schools in targeting strategies consistent with the components of the MUSIC model in hopes that these strategies will alter the documented decline in motivation for the sciences, which may serve to help assuage our national debt in science professionals. For example, a recent study of an afterschool science and engineering program that incorporated elements of the MUSIC model indicated that those students who participated in two phases of the extracurricular program maintained their motivation for science over time, whereas their peers tended to follow the expected decline (Chittum et al., under review).



## Relationships to Science Identification

Our findings provide evidence to support some of the relationships within the domain identification model (Osborne & Jones, 2011). As predicted by theoretical and empirical evidence (Jones, Ruff, et al., 2015; Jones et al., 2014; Osborne & Jones, 2011), students' science class perceptions of the MUSIC model components were statistically related to their science identification; and, in addition, science identification was statistically related to science effort, course intentions, and career goals. The fact that the higher-numbered motivation profiles were more strongly related to higher levels of science identification than the lower-numbered motivation profiles provides further evidence that these variables are positively correlated.

## Limitations

This study serves as a proof of concept and, thus, is the first step in investigating class-related motivation profiles of pre-high school students in science considering science class perceptions related to empowerment, usefulness, success, interest, and caring. More research will be needed to examine teacher effects, motivation profiles in different contexts (e.g., domains, grade levels, schools), effects of designing interventions for motivation profiles, and further understanding the implications of this person-centered approach. We are unsure how much of the students' perceptions about science class can be attributed to personal traits and how much can be attributed to the class—another area for future research. Also, collecting qualitative data may help researchers to better interpret and explain these profiles.

In general, studies that use self-report measures are inherently subject to validity threats due to inaccurate assessment of personal beliefs, misunderstandings, and examining perceptions that may be novel to participants and thus lack prior thought. As all students were enrolled in two schools in one rural area, the results of this study may not be generalizable to other students who vary significantly from the participants in this study. In cluster analysis especially, which is exploratory in nature, cluster membership and structure can vary depending on the context; therefore, we caution against overgeneralizing across samples, environments, and domains (Vansteenkiste et al., 2009). Similarly, the MUSIC model summarizes important teaching strategies such that they can be communicated to, and utilized by, practicing educators. As in all situations in which concepts are summarized, important information can be lost; thus, distinctions among the strategies and theories from which the MUSIC model stemmed may be lost or made somewhat unclear. As noted previously, it is also possible that our measures of the MUSIC model components did not assess the range of perceptions that are possible within each component. Nonetheless, we believe that this study represents an important contribution in understanding how multidimensional motivation perceptions affect students' outcomes in science classes.

## Conclusion

We identified five different class-related motivation profiles that illustrate complex patterns in students' perceptions about science

class and appear to manifest differently among individuals. The five profiles formed consistent patterns in students' self-reported effort in science class, and aligned as expected with theoretically and empirically correlated variables such as science identification, course intentions, and career goals. These findings indicate that students' perceptions of the MUSIC model components in science class and their levels of science identification are important to consider when trying to motivate students to both engage in their current science class and consider a science-related career in the future.

The fact that students' perceptions varied across the five MUSIC model components contributes to the literature in a few ways. First, it demonstrates that motivation in science class is a multidimensional construct that comprises different facets that can be assessed quickly with a paper-and-pencil questionnaire. Theoretically, this raises questions about how these science class perceptions (i.e., the MUSIC model components), drawn from multiple motivation theories, can be integrated into a more comprehensive theory of students' motivation in class. Further research is needed to better understand how the five MUSIC model components work together to influence students' motivation in a class and how these perceptions then affect their domain identification and goals. Practically, because the MUSIC model components are important to students' motivation, teachers should assess these components (either formally or informally) and then identify strategies to help students develop positive perceptions of their classes. The profiles identified in this study may help educators to intentionally design instruction for students with similar class-related motivation profiles, rather than adhere to the difficult and often unrealistic task of targeting each student's individual complex needs. Furthermore, these findings underscore the importance of the teacher's instructional decisions in impacting students' motivation-related perceptions. Thus, using these profiles, teachers may be able to more purposefully affect students' motivation in science classrooms and increase the likelihood that more students will engage in science class and consider science-related careers.

## References

- Ainsworth, M. D. S. (1973). The development of infant-mother attachment. In B. Caldwell & H. Ricciuti (Eds.), *Review of child development research* (Vol. 3, pp. 1–94). Chicago, IL: University of Chicago Press.
- Aschbacher, P. R., Ing, M., & Tsai, S. M. (2014). Is science me? Exploring middle school students' STE-M career. *Journal of Science Education and Technology*, 23, 735–743. <http://dx.doi.org/10.1007/s10956-014-9504-x>
- Asendorpf, J. B., Borkenau, P., Ostendorf, F., & van Aken, M. A. G. (2001). Carving personality description at its joints: Confirmation of three replicable personality prototypes for both children and adults. *European Journal of Personality*, 15, 169–198. <http://dx.doi.org/10.1002/per.408>
- Assor, A., Kaplan, H., & Roth, G. (2002). Choice is good, but relevance is excellent: Autonomy-enhancing and suppressing teacher behaviours predicting students' engagement in schoolwork. *British Journal of Educational Psychology*, 72, 261–278. <http://dx.doi.org/10.1348/000709902158883>
- Bacher, J. (2002). *Cluster analysis*. Retrieved from <http://www.clusteranalyse.net/sonstiges/zaspringseminar2002/lecturenotes.pdf>



- Bacher, J., Wenzig, K., & Vogler, M. (2004). *SPSS twostep cluster: A first evaluation*. Retrieved from [http://opus4.kobv.de/opus4-fau/files/74/a\\_04-02.pdf](http://opus4.kobv.de/opus4-fau/files/74/a_04-02.pdf)
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.
- Barak, M., Ashkar, T., & Dori, Y. J. (2011). Learning science via animated movies: Its effect on students' thinking and motivation. *Computers & Education*, 56, 839–846. <http://dx.doi.org/10.1016/j.compedu.2010.10.025>
- Bartholomew, D. J., Steele, F., Galbraith, J., & Moustaki, I. (2008). *Analysis of multivariate social science data* (2nd ed.). Boca Raton, FL: Taylor & Francis.
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117, 497–529. <http://dx.doi.org/10.1037/0033-2909.117.3.497>
- Berger, R., & Hänze, M. (2009). Comparison of two small-group learning methods in 12th-grade physics classes focusing on intrinsic motivation and academic performance. *International Journal of Science Education*, 31, 1511–1527. <http://dx.doi.org/10.1080/09500690802116289>
- Bergin, C., & Bergin, D. (2009). Attachment in the classroom. *Educational Psychology Review*, 21, 141–170. <http://dx.doi.org/10.1007/s10648-009-9104-0>
- Bergin, D. A. (1999). Influences on classroom interest. *Educational Psychologist*, 34, 87–98. [http://dx.doi.org/10.1207/s15326985ep3402\\_2](http://dx.doi.org/10.1207/s15326985ep3402_2)
- Bergman, L. R. (2001). A person approach in research on adolescence: Some methodological challenges. *Journal of Adolescent Research*, 16, 28–53. <http://dx.doi.org/10.1177/0743558401161004>
- Bong, M., Lee, S. K., & Woo, Y. (2015). The roles of interest and self-efficacy in the decision to pursue mathematics and science. In K. A. Renninger, M. Nieswandt, & S. Hidi (Eds.), *Interest in mathematics and science learning* (pp. 33–48). Washington, DC: American Educational Research Association. [http://dx.doi.org/10.3102/978-0-935302-42-4\\_2](http://dx.doi.org/10.3102/978-0-935302-42-4_2)
- Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, 15, 1–40. <http://dx.doi.org/10.1023/A:1021302408382>
- Bowers, A. J., & Sprott, R. (2012). Examining the multiple trajectories associated with dropping out of high school: A growth mixture model analysis. *The Journal of Educational Research*, 105, 176–195. <http://dx.doi.org/10.1080/00220671.2011.552075>
- Bowlby, J. (1969). *Attachment* (Vol. I). New York, NY: Basic.
- Breckenridge, J. N. (2000). Validating cluster analysis: Consistent replication and symmetry. *Multivariate Behavioral Research*, 35, 261–285. [http://dx.doi.org/10.1207/S15327906MBR3502\\_5](http://dx.doi.org/10.1207/S15327906MBR3502_5)
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 445–455). Newbury Park, CA: Sage.
- Bulunuz, M., & Jarrett, O. S. (2015). Play as an aspect of interest development in science. In K. A. Renninger, M. Nieswandt, & S. Hidi (Eds.), *Interest in mathematics and science learning* (pp. 153–171). Washington, DC: American Educational Research Association. [http://dx.doi.org/10.3102/978-0-935302-42-4\\_9](http://dx.doi.org/10.3102/978-0-935302-42-4_9)
- Burns, R., & Burns, R. (2008). *Business research methods and statistics using SPSS*. Thousand Oaks, CA: Sage.
- Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.
- Chen, J. A. (2012). Implicit theories, epistemic beliefs, and science motivation: A person-centered approach. *Learning and Individual Differences*, 22, 724–735. <http://dx.doi.org/10.1016/j.lindif.2012.07.013>
- Cheung, D. (2015). The combined effects of classroom teaching and learning strategy use on students' chemistry self-efficacy. *Research in Science Education*, 45, 101–116. <http://dx.doi.org/10.1007/s11165-014-9415-0>
- Chittum, J. R., Jones, B. D., Akalin, S., & Schram, A. B. (under review). The effects of an afterschool STEM program on students' motivation and engagement.
- Chittum, J. R., McConnell, K. D., & Sible, J. (in press). Undergraduate students' perceptions of motivation in a SCALE-UP Cancer Biology course: A case study. *Journal on Excellence in College Teaching*.
- Conley, A. M. (2012). Patterns of motivation beliefs: Combining achievement goal and expectancy-value perspectives. *Journal of Educational Psychology*, 104, 32–47. <http://dx.doi.org/10.1037/a0026042>
- Cox, A. E., & Whaley, D. E. (2004). The influence of task value, expectancies for success, and identity on athletes' achievement behaviors. *Journal of Applied Sport Psychology*, 16, 103–117. <http://dx.doi.org/10.1080/10413200490437930>
- Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York, NY: Harper & Row.
- Daniels, L. M., Haynes, T. L., Stupnisky, R. H., Perry, R. P., Newall, N. E., & Pekrun, R. (2008). Individual differences in achievement goals: A longitudinal study of cognitive, emotional, and achievement outcomes. *Contemporary Educational Psychology*, 33, 584–608. <http://dx.doi.org/10.1016/j.cedpsych.2007.08.002>
- DeBacker, T. K., & Nelson, R. M. (2000). Motivation to learn science: Differences related to gender, class type, and ability. *The Journal of Educational Research*, 93, 245–254. <http://dx.doi.org/10.1080/00220670009598713>
- deCharms, R. (1968). *Personal causation: The internal affective determinants of behavior*. New York, NY: Academic Press.
- Deci, E. L. (1975). *Intrinsic motivation*. New York, NY: Plenum Press. <http://dx.doi.org/10.1007/978-1-4613-4446-9>
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum Press. <http://dx.doi.org/10.1007/978-1-4899-2271-7>
- Deci, E. L., & Ryan, R. M. (1991). A motivational approach to self: Integration in personality. In R. Dientsbier (Ed.), *Nebraska Symposium on Motivation* (Vol. 38, pp. 237–288). Lincoln, NE: University of Nebraska Press.
- Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11, 227–268. [http://dx.doi.org/10.1207/S15327965PLI1104\\_01](http://dx.doi.org/10.1207/S15327965PLI1104_01)
- Deci, E. L., & Ryan, R. M. (2012). Motivation, personality, and development within embedded social contexts: An overview of self-determination theory. In R. M. Ryan (Ed.), *The Oxford handbook of human motivation* (pp. 85–107). New York, NY: Oxford University Press. <http://dx.doi.org/10.1093/oxfordhb/9780195399820.013.0006>
- De Volder, M., & Lens, W. (1982). Academic achievement and future time perspective as a cognitive-motivational concept. *Journal of Personality and Social Psychology*, 42, 566–571. <http://dx.doi.org/10.1037/0022-3514.42.3.566>
- Dotterer, A. M., & Lowe, K. (2011). Classroom context, school engagement, and academic achievement in early adolescence. *Journal of Youth and Adolescence*, 40, 1649–1660. <http://dx.doi.org/10.1007/s10964-011-9647-5>
- Eccles, J. S. (2007). Where are all the women? Gender differences in participation in physical science and engineering. In S. J. Ceci & W. M. Williams (Eds.), *Why aren't more women in science? Top researchers debate the evidence* (pp. 199–210). Washington, DC: American Psychological Association. <http://dx.doi.org/10.1037/11546-016>
- Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: Freeman.
- Eccles, J. S., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin*, 21, 215–225. <http://dx.doi.org/10.1177/0146167295213003>



- Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development*, 64, 830–847. <http://dx.doi.org/10.2307/1131221>
- Eccles, J. S., Wigfield, A., & Schiefele, U. (1998). Motivation to succeed. In W. Damon & N. Eisenberg (Eds.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (5th ed., pp. 1017–1095). Hoboken, NJ: Wiley.
- Egan, O. (1984). Cluster analysis in educational research. *British Educational Research Journal*, 10, 145–153. <http://dx.doi.org/10.1080/0141192840100203>
- Elliot, A. J., & Dweck, C. S. (2005). Competence and motivation: Competence as the core of achievement motivation. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 3–12). New York, NY: Guilford Press.
- Fischer, F., Schult, J., & Hell, B. (2013). Sex differences in secondary school success: Why female students perform better. *European Journal of Psychology of Education*, 28, 529–543. <http://dx.doi.org/10.1007/s10212-012-0127-4>
- Flleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York, NY: Wiley.
- Fortus, D., & Vedder-Weiss, D. (2014). Measuring students' continuing motivation for science learning. *Journal of Research in Science Teaching*, 51, 497–522. <http://dx.doi.org/10.1002/tea.21136>
- Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41, 578–588. <http://dx.doi.org/10.1093/comjnl/41.8.578>
- Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611–631. <http://dx.doi.org/10.1198/016214502760047131>
- Galbraith, J. K., & Jiaqing, L. (1999). *Cluster and discriminant analysis on time-series as a research tool*. Retrieved from [http://utip.gov.utexas.edu/papers/utip\\_06.pdf](http://utip.gov.utexas.edu/papers/utip_06.pdf)
- Geiser, C., Lehmann, W., & Eid, M. (2006). Separating “rotators” from “nonrotators” in the mental rotations test: A multigroup latent class analysis. *Multivariate Behavioral Research*, 41, 261–293. [http://dx.doi.org/10.1207/s15327906mbr4103\\_2](http://dx.doi.org/10.1207/s15327906mbr4103_2)
- Gendolla, G. H. E., Wright, R. A., & Richter, M. (2012). Effort intensity: Some insights from the cardiovascular system. In R. M. Ryan (Ed.), *The Oxford handbook of human motivation* (pp. 420–440). New York, NY: Oxford University Press. <http://dx.doi.org/10.1093/oxfordhb/9780195399820.013.0024>
- Greene, T. K., DeBacker, B. R., & Krows, A. J. (1999). Goals, values, and beliefs as predictors of achievement effort in high school mathematics classes. *Sex Roles*, 40(5/6), 421–458. <http://dx.doi.org/10.1023/A:1018871610174>
- Hafen, C. A., Allen, J. P., Mikami, A. Y., Gregory, A., Hamre, B., & Pianta, R. C. (2012). The pivotal role of adolescent autonomy in secondary school classrooms. *Journal of Youth and Adolescence*, 41, 245–255. <http://dx.doi.org/10.1007/s10964-011-9739-2>
- Hale, R. L., & Glassman, S. S. (1986). Using computers to construct, evaluate, and apply actuarial systems of classification: A methodology and example. *Computers in Human Behavior*, 2, 195–213. [http://dx.doi.org/10.1016/0747-5632\(86\)90003-8](http://dx.doi.org/10.1016/0747-5632(86)90003-8)
- Hartwell, M., & Kaplan, A. (2014, April). *A multidimensional analysis of context-specific perceived relevance among students with differing expectancy-value profiles*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.
- Hayenga, A. O., & Corpus, J. H. (2010). Profiles of intrinsic and extrinsic motivations: A person-centered approach to motivation and achievement in middle school. *Motivation and Emotion*, 34, 371–383. <http://dx.doi.org/10.1007/s11031-010-9181-x>
- Hidi, S., & Harackiewicz, J. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, 70, 151–179. <http://dx.doi.org/10.3102/00346543070002151>
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 41, 111–127. [http://dx.doi.org/10.1207/s15326985ep4102\\_4](http://dx.doi.org/10.1207/s15326985ep4102_4)
- Hidi, S., Renninger, K. A., & Nieswandt, M. (2015). Emerging issues and themes in addressing interest in learning mathematics and science. In K. A. Renninger, M. Nieswandt, & S. Hidi (Eds.), *Interest in mathematics and science learning* (pp. 385–396). Washington, DC: American Educational Research Association. [http://dx.doi.org/10.3102/978-0-935302-42-4\\_Cncln](http://dx.doi.org/10.3102/978-0-935302-42-4_Cncln)
- Hoffmann, L. (2002). Promoting girls' interest and achievement in physics classes for beginners. *Learning and Instruction*, 12, 447–465. [http://dx.doi.org/10.1016/S0959-4752\(01\)00010-X](http://dx.doi.org/10.1016/S0959-4752(01)00010-X)
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. <http://dx.doi.org/10.1080/10705519909540118>
- Huberty, C. J., Jordan, E. M., & Brandt, W. C. (2005). Cluster analysis in higher education research. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. XX, pp. 437–457). Dordrecht, the Netherlands: Springer. [http://dx.doi.org/10.1007/1-4020-3279-X\\_8](http://dx.doi.org/10.1007/1-4020-3279-X_8)
- Hulleman, C. S., Durik, A. M., Schweigert, S. A., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology*, 100, 398–416. <http://dx.doi.org/10.1037/0022-0663.100.2.398>
- Ireson, J., & Hallam, S. (2005). Pupils' liking for school: Ability grouping, self-concept and perceptions of teaching. *British Journal of Educational Psychology*, 75, 297–311. <http://dx.doi.org/10.1348/000709904X24762>
- Jacobs, J. E., Finken, L. L., Griffin, N. L., & Wright, J. D. (1998). The career plans of science-talented rural adolescent girls. *American Educational Research Journal*, 35, 681–704. <http://dx.doi.org/10.3102/00028312035004681>
- Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development*, 73, 509–527. <http://dx.doi.org/10.1111/1467-8624.00421>
- Janssen, A. B., & Geiser, C. (2010). On the relationship between solution strategies in two mental rotation tasks. *Learning and Individual Differences*, 20, 473–478. <http://dx.doi.org/10.1016/j.lindif.2010.03.002>
- Jen, T.-H., Lee, C.-D., Chien, C.-L., Hsu, Y.-S., & Chen, K.-M. (2013). Perceived social relationships and science learning outcomes for Taiwanese eighth graders: Structural equation modeling with a complex sampling consideration. *International Journal of Science and Mathematics Education*, 11, 575–600. <http://dx.doi.org/10.1007/s10763-012-9355-y>
- Jiang, M. F., Tseng, S. S., & Su, C. M. (2001). Two-phase clustering process for outliers detection. *Pattern Recognition Letters*, 22, 691–700. [http://dx.doi.org/10.1016/S0167-8655\(00\)00131-8](http://dx.doi.org/10.1016/S0167-8655(00)00131-8)
- Jones, B. D. (2009). Motivating students to engage in learning: The MUSIC model of academic motivation. *International Journal on Teaching and Learning in Higher Education*, 21, 272–285. Retrieved from <http://www.isetl.org/ijtlhe/pdf/IJTLHE774.pdf>
- Jones, B. D. (2010). An examination of motivation model components in face-to-face and online instruction. *Electronic Journal of Research in Educational Psychology*, 8, 915–944.
- Jones, B. D. (2015). *Motivating students by design: Practical strategies for professors*. Charleston, SC: CreateSpace.
- Jones, B. D. (2016a). Teaching motivation strategies using the MUSIC® Model of Motivation as a conceptual framework. In M. C. Smith & N. DeFrates-Densch (Eds.), *Challenges and innovations in educational psychology teaching and learning* (pp. 123–136). Charlotte, NC: Information Age.



- Jones, B. D. (2016b). *User guide for assessing the components of the MUSIC Model of Academic Motivation*. Retrieved from <http://www.theMUSICmodel.com>
- Jones, B. D., Chittum, J. R., Akalin, S., Schram, A. B., Fink, J., Schnittka, C., . . . Brandt, C. (2015). Elements of design-based science activities that affect students' motivation. *School Science and Mathematics, 115*, 404–415. <http://dx.doi.org/10.1111/ssm.12143>
- Jones, B. D., Epler, C. M., Mokri, P., Bryant, L. H., & Paretti, M. C. (2013). The effects of a collaborative problem-based learning experience on students' motivation in engineering capstone courses. *Interdisciplinary Journal of Problem-based Learning, 7*, 34–71. <http://dx.doi.org/10.7771/1541-5015.1344>
- Jones, B. D., Li, M., & Cruz, J. M. (2017). A cross-cultural validation of the MUSIC® Model of Academic Motivation Inventory: Evidence from Chinese- and Spanish-speaking university students. *International Journal of Educational Psychology, 6*, 366–385. <http://dx.doi.org/10.17583/ijep.2017.2357>
- Jones, B. D., Osborne, J. W., Paretti, M. C., & Matusovich, H. M. (2014). Relationships among students' perceptions of a first-year engineering design course and their engineering identification, motivational beliefs, course effort, and academic outcomes. *International Journal of Engineering Education, 30*, 1340–1356.
- Jones, B. D., Paretti, M. C., Hein, S. F., & Knott, T. W. (2010). An analysis of motivation constructs with first-year engineering students: Relationships among expectancies, values, achievement, and career plans. *Journal of Engineering Education, 99*, 319–336. <http://dx.doi.org/10.1002/j.2168-9830.2010.tb01066.x>
- Jones, B. D., Ruff, C., & Osborne, J. W. (2015). Fostering students' identification with mathematics and science. In K. A. Renninger, M. Nieswandt, & S. Hidi (Eds.), *Interest in mathematics and science learning* (pp. 331–352). Washington, DC: American Educational Research Association. [http://dx.doi.org/10.3102/978-0-935302-42-4\\_19](http://dx.doi.org/10.3102/978-0-935302-42-4_19)
- Jones, B. D., Ruff, C., Snyder, J. D., Petrich, B., & Koonce, C. (2012). The effects of mind mapping activities on students' motivation. *International Journal for the Scholarship of Teaching and Learning, 6*, 1–21. <http://dx.doi.org/10.20429/ijstl.2012.060105>
- Jones, B. D., & Sigmon, M. L. (2016). Validation evidence for the elementary school version of the MUSIC® Model of Academic Motivation Inventory. *Electronic Journal of Research in Educational Psychology, 14*, 155–174. <http://dx.doi.org/10.14204/ejrep.38.15081>
- Jones, B. D., & Skaggs, G. E. (2016). Measuring students' motivation: Validity evidence for the MUSIC Model of Academic Motivation Inventory. *International Journal for the Scholarship of Teaching and Learning, 10*. Retrieved from <http://digitalcommons.georgiasouthern.edu/ijstl/vol10/iss1/7>
- Jones, B. D., Tendhar, C., & Paretti, M. C. (2016). The effects of students' course perceptions on their domain identification, motivational beliefs, and goals. *Journal of Career Development, 433*, 83–397.
- Jones, B. D., Watson, J. M., Rakes, L., & Akalin, S. (2012). Factors that impact students' motivation in an online course: Using the MUSIC Model of Academic Motivation. *Journal of Teaching and Learning with Technology, 1*, 42–58.
- Jones, B. D., & Wilkins, J. L. M. (2013). Testing the MUSIC Model of Academic Motivation through confirmatory factor analysis. *Educational Psychology, 33*, 482–503. <http://dx.doi.org/10.1080/01443410.2013.785044>
- Jung, J.-S., Owusu-Antwi, E. B., & An, J.-H. (2006). Analytical procedures for evaluating factors that affect joint faulting for jointed plain concrete pavements using the long-term pavements performance database. *Canadian Journal of Civil Engineering, 33*, 1279–1286. <http://dx.doi.org/10.1139/106-072>
- Kaplan, A., Katz, I., & Flum, H. (2012). Motivation theory in educational practice: Knowledge claims, challenges, and future directions. In K. R. Harris, S. G. Graham, & T. Urdan (Eds.), *APA educational psychology handbook: Vol. 2. Individual differences, cultural considerations, and contextual factors in educational psychology* (pp. 165–194). Washington, DC: American Psychological Association. <http://dx.doi.org/10.1037/13274-007>
- Kaplan, A., Katz, I., & Flum, H. (2014, April). Practice-relevant motivational research: Do we need a different approach? In L. Corno (Chair), *Practice-relevant motivational research: Do we need a different approach?* Symposium conducted at the annual meeting of the American Educational Research Association, Philadelphia, PA.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York, NY: Wiley. <http://dx.doi.org/10.1002/9780470316801>
- Kline, R. B. (2005). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. <http://dx.doi.org/10.2307/2529310>
- Lange, T., Roth, V., Braun, M. L., & Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural Computation, 16*, 1299–1323. <http://dx.doi.org/10.1162/089976604773717621>
- Lee, J. D. (2002). More than ability: Gender and personal relationships influence science and technology involvement. *Sociology of Education, 75*, 349–373. <http://dx.doi.org/10.2307/3090283>
- Lee, W. C., Kajfez, R. L., & Matusovich, H. M. (2013). Motivating engineering students: Evaluating an engineering student support center with the MUSIC model of academic motivation. *Journal of Women and Minorities in Science and Engineering, 19*, 245–271. <http://dx.doi.org/10.1615/JWomenMinorScienEng.2013006747>
- Legault, L., Green-Demers, I., & Pelletier, L. (2006). Why do high school students lack motivation in the classroom? Toward an understanding of academic amotivation and the role of social support. *Journal of Educational Psychology, 98*, 567–582. <http://dx.doi.org/10.1037/0022-0663.98.3.567>
- Maltese, A. V., & Harsh, J. A. (2015). Students' pathways of entry into STEM. In K. A. Renninger, M. Nieswandt, & S. Hidi (Eds.), *Interest in mathematics and science learning* (pp. 203–223). Washington, DC: American Educational Research Association. [http://dx.doi.org/10.3102/978-0-935302-42-4\\_12](http://dx.doi.org/10.3102/978-0-935302-42-4_12)
- Maltese, A. V., & Tai, R. H. (2010). Eyeballs in the fridge: Sources of early interest in science. *International Journal of Science Education, 32*, 669–685. <http://dx.doi.org/10.1080/09500690902792385>
- McGinley, J., & Jones, B. D. (2014). A brief instructional intervention to increase students' motivation on the first day of class. *Teaching of Psychology, 41*, 158–162. <http://dx.doi.org/10.1177/0098628314530350>
- Meece, J. L., Glienke, B. B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology, 44*, 351–373. <http://dx.doi.org/10.1016/j.jsp.2006.04.004>
- Meece, J. L., & Holt, K. (1993). A pattern analysis of students' achievement goals. *Journal of Educational Psychology, 85*, 582–590. <http://dx.doi.org/10.1037/0022-0663.85.4.582>
- Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology, 82*, 60–70. <http://dx.doi.org/10.1037/0022-0663.82.1.60>
- Midgley, C., & Feldlaufer, H. (1987). Students' and teachers' decision making fit before and after the transition to junior high school. *The Journal of Early Adolescence, 7*, 225–241. <http://dx.doi.org/10.1177/0272431687072009>
- Miller, R. B., Greene, B. A., Montalvo, G. P., Ravindran, B., & Nichols, J. D. (1996). Engagement in academic work: The role of learning goals, future consequences, pleasing others, and perceived ability. *Contemporary Educational Psychology, 21*, 388–422. <http://dx.doi.org/10.1006/ceps.1996.0028>



- Mohamed, H. E., Soliman, M. H., & Jones, B. D. (2013). A cross-cultural validation of the MUSIC Model of Academic Motivation and its associated inventory among Egyptian university students. *Journal of Counseling Quarterly Journal*, 36, 2–14.
- Mooi, E., & Sarstedt, M. (2011). *A concise guide to market research*. Berlin, Germany: Springer-Verlag. <http://dx.doi.org/10.1007/978-3-642-12541-6>
- Mora, C. E., Anorbe-Diaz, B., Gonzalez-Marrero, A. M., Martin-Gutierrez, J., & Jones, B. D. (in press). Motivational aspects when introducing problem-based learning in engineering education. *International Journal of Engineering Education*.
- National Academy of Sciences (NAS). (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Washington, DC: National Academies Press. Retrieved from <http://www.sandia.gov/NINE/documents/RisingAbove.pdf>
- NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: National Academies Press.
- Nieswandt, M., & Shanahan, M.-C. (2008). "I just want the credit!"—Perceived instrumentality as the main characteristic of boys' motivation in a grade 11 science course. *Research in Science Education*, 38, 3–29. <http://dx.doi.org/10.1007/s11165-007-9037-x>
- Noddings, M. (1992). *The challenge to care in schools: An alternative approach to education*. New York, NY: Teachers College Press.
- Norušis, M. J. (2011). *IBM SPSS statistics 19 statistical procedures companion*. Upper Saddle River, NJ: Prentice Hall.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535–569. <http://dx.doi.org/10.1080/10705510701575396>
- Osborne, J. W. (1997). Identification with academics and academic success among community college students. *Community College Review*, 25, 59–67. <http://dx.doi.org/10.1177/009155219702500105>
- Osborne, J. W., & Jones, B. D. (2011). Identification with academics and motivation to achieve in school: How the structure of the self influences academic outcomes. *Educational Psychology Review*, 23, 131–158. <http://dx.doi.org/10.1007/s10648-011-9151-1>
- Osborne, J. W., & Rausch, J. L. (2001, April). Identification with academics and academic outcomes in secondary students. Paper presented at the annual meeting of the American Education Research Association, Seattle, WA.
- Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, 25, 1049–1079. <http://dx.doi.org/10.1080/0950069032000032199>
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, 66, 543–578. <http://dx.doi.org/10.3102/00346543066004543>
- Parkes, K., Jones, B. D., & Wilkins, J. (2015). Assessing music students' motivation using the MUSIC Model of Academic Motivation Inventory. *UPDATE: Applications of Research in Music Education*. Advance online publication. <http://dx.doi.org/10.1177/8755123315620835>
- Plant, R. W., & Ryan, R. M. (1985). Intrinsic motivation and the effects of self-consciousness, self-awareness, and ego-involvement: An investigation of internally-controlling styles. *Journal of Personality*, 53, 435–449. <http://dx.doi.org/10.1111/j.1467-6494.1985.tb00375.x>
- President's Council of Advisors on Science and Technology (PCAST). (2010). *Report to the President: Prepare and inspire: K-12 education in science, technology, engineering, and math (STEM) for America's future*. Retrieved from <http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-stemed-report.pdf>
- President's Council of Advisors on Science and Technology (PCAST). (2012). *Report to the president: Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Retrieved from [http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-executive-report-final\\_2-13-12.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-executive-report-final_2-13-12.pdf)
- Ratelle, C. F., Guay, F., Vallerand, R. J., Larose, S., & Senécal, C. (2007). Autonomous, controlled, and amotivated types of academic motivation: A person-oriented analysis. *Journal of Educational Psychology*, 99, 734–746. <http://dx.doi.org/10.1037/0022-0663.99.4.734>
- Reeve, J. (1989). The interest-enjoyment distinction in intrinsic motivation. *Motivation and Emotion*, 13, 83–103. <http://dx.doi.org/10.1007/BF00992956>
- Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion*, 28, 147–169. <http://dx.doi.org/10.1023/B:MOEM.0000032312.95499.6f>
- Reeve, J., Nix, G., & Hamm, D. (2003). Testing models of the experience of self-determination in intrinsic motivation and the conundrum of choice. *Journal of Educational Psychology*, 95, 375–392. <http://dx.doi.org/10.1037/0022-0663.95.2.375>
- Reilly, C., Wang, C., & Rutherford, M. (2005). A rapid method for the comparison of cluster analyses. *Statistica Sinica*, 15, 19–33. Retrieved from <http://www3.stat.sinica.edu.tw/statistica/oldpdf/A15n12.pdf>
- Renninger, K. A., Hidi, S., & Krapp, A. (Eds.). (1992). *The role of interest in learning and development*. Hillsdale, NJ: Erlbaum.
- Renninger, K. A., Nieswandt, M., & Hidi, S. (Eds.). (2015). *Interest in mathematics and science learning*. Washington, DC: American Educational Research Association. <http://dx.doi.org/10.3102/978-0-935302-42-4>
- Reynolds, B., Mehalik, M. M., Lovell, M. R., & Schunn, C. D. (2009). Increasing student awareness of and interest in engineering as a career option through design-based learning. *International Journal of Engineering Education*, 25, 788–798.
- Rosen, Y. (2009). The effects of an animation-based online learning environment on transfer of knowledge and on motivation for science and technology learning. *Journal of Educational Computing Research*, 40, 451–467. <http://dx.doi.org/10.2190/EC.40.4.d>
- Rudasill, K. M., & Callahan, C. M. (2010). Academic self-perceptions of ability and course planning among academically advanced students. *Journal of Advanced Academics*, 21, 300–329. <http://dx.doi.org/10.1177/1932202X1002100206>
- Ryan, R. M. (1995). Psychological needs and the facilitation of integrative processes. *Journal of Personality*, 63, 397–427. <http://dx.doi.org/10.1111/j.1467-6494.1995.tb00501.x>
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78. <http://dx.doi.org/10.1037/0003-066X.55.1.68>
- Schmader, T., Major, B., & Gramzow, R. H. (2001). Coping with ethnic stereotypes in the academic domain: Perceived injustice and psychological disengagement. *Journal of Social Issues*, 57, 93–111. <http://dx.doi.org/10.1111/0022-4537.00203>
- Schram, A. B., & Jones, B. D. (2016). A cross-cultural adaptation and validation of the Icelandic version of the MUSIC Model of Academic Motivation Inventory. *Icelandic Journal of Education*, 25, 159–181.
- Schraw, G., & Lehman, S. (2001). Situational interest: A review of the literature and directions for future research. *Educational Psychology Review*, 13, 23–52. <http://dx.doi.org/10.1023/A:1009004801455>
- Schunk, D. H. (1995). Self-efficacy and education and instruction. In J. E. Maddux (Ed.), *Self-efficacy, adaptation, and adjustment: Theory, research, and application* (pp. 281–303). New York, NY: Plenum Press. [http://dx.doi.org/10.1007/978-1-4419-6868-5\\_10](http://dx.doi.org/10.1007/978-1-4419-6868-5_10)
- Schunk, D. H., & Pajares, F. (2005). Competence perceptions and academic functioning. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 141–163). New York, NY: Guilford Press.



- Schwinger, M., Steinmayr, R., & Spinath, B. (2012). Not all roads lead to Rome—Comparing different types of motivational regulation profiles. *Learning and Individual Differences*, 22, 269–279. <http://dx.doi.org/10.1016/j.lindif.2011.12.006>
- Scogin, S. C., & Stuessy, C. L. (2015). Encouraging greater student inquiry engagement in science through motivational support by online scientist-mentors. *Science Education*, 99, 312–349. <http://dx.doi.org/10.1002/sce.21145>
- Shell, D. F., & Husman, J. (2008). Control, motivation, affect, and strategic self-regulation in the college classroom: A multidimensional phenomenon. *Journal of Educational Psychology*, 100, 443–459. <http://dx.doi.org/10.1037/0022-0663.100.2.443>
- Shell, D. F., & Soh, L. K. (2013). Profiles of motivated self-regulation in college computer science courses: Differences in major versus required non-major courses. *Journal of Science Education and Technology*, 22, 899–913. <http://dx.doi.org/10.1007/s10956-013-9437-9>
- Simons, J., Vansteenkiste, M., Lens, W., & Lacante, M. (2004). Placing motivation and future time perspective theory in a temporal perspective. *Educational Psychology Review*, 16, 121–139. <http://dx.doi.org/10.1023/B:EDPR.0000026609.94841.2f>
- Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology*, 42, 70–83. <http://dx.doi.org/10.1037/0012-1649.42.1.70>
- Simpson, R. D., & Oliver, J. S. (1990). A summary of major influences on attitude toward and achievement in science among adolescent students. *Science Education*, 74, 1–18. <http://dx.doi.org/10.1002/sce.3730740102>
- Smith, J. (2012, May 29). The 10 hardest jobs to fill in America. *Forbes*. Retrieved from <http://www.forbes.com/sites/jacquelynsmith/2012/05/29/the-10-hardest-jobs-to-fill-in-america-2/>
- Spiegel, A. N., McQuillan, J., Halpin, P., Matuk, C., & Diamond, J. (2013). Engaging teenagers with science through comics. *Research in Science Education*, 43, 2309–2326. <http://dx.doi.org/10.1007/s11165-013-9358-x>
- Stake, J. E., & Nickens, S. D. (2005). Adolescent girls' and boys' science peer relationships and perceptions of the possible self as scientist. *Sex Roles*, 52, 1–11. <http://dx.doi.org/10.1007/s11199-005-1189-4>
- Steinmayr, R., & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality*, 22, 185–209. <http://dx.doi.org/10.1002/per.676>
- Tai, R. H., Qi Liu, C., Maltese, A. V., & Fan, X. (2006). Career choice. Planning early for careers in science. *Science*, 312, 1143–1144. <http://dx.doi.org/10.1126/science.1128690>
- Tsai, Y., Kunter, M., Lüdtke, O., Trautwein, U., & Ryan, R. M. (2008). What makes lessons interesting? The role of situational and individual factors in three school subjects. *Journal of Educational Psychology*, 100, 460–472. <http://dx.doi.org/10.1037/0022-0663.100.2.460>
- Tuominen-Soini, H., Salmela-Aro, K., & Niemivirta, M. (2011). Stability and change in achievement goal orientations: A person-centered approach. *Contemporary Educational Psychology*, 36, 82–100. <http://dx.doi.org/10.1016/j.cedpsych.2010.08.002>
- Turner, J. C., Christensen, A., Kackar-Cam, H. Z., Trucano, M., & Fulmer, S. M. (2014). Enhancing students' engagement: Report of a 3-year intervention with middle school teachers. *American Educational Research Journal*, 51, 1195–1226. <http://dx.doi.org/10.3102/0002831214532515>
- Turner, J. C., & Meyer, D. K. (2000). Studying and understanding the instructional contexts of classrooms: Using our past to forge our future. *Educational Psychologist*, 35, 69–85. [http://dx.doi.org/10.1207/S15326985EP3502\\_2](http://dx.doi.org/10.1207/S15326985EP3502_2)
- Turner, J. C., Thorpe, P. K., & Meyer, D. K. (1998). Students' reports of motivation and negative affect: A theoretical and empirical analysis. *Journal of Educational Psychology*, 90, 758–771. <http://dx.doi.org/10.1037/0022-0663.90.4.758>
- Urdu, T., & Schoenfelder, E. (2006). Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs. *Journal of School Psychology*, 44, 331–349. <http://dx.doi.org/10.1016/j.jsp.2006.04.003>
- Vansteenkiste, M., Lens, W., & Deci, E. L. (2006). Intrinsic versus extrinsic goal contents in self-determination theory: Another look at the quality of academic motivation. *Educational Psychologist*, 41, 19–31. [http://dx.doi.org/10.1207/s15326985ep4101\\_4](http://dx.doi.org/10.1207/s15326985ep4101_4)
- Vansteenkiste, M., & Mouratidis, A. (2016). Emerging trends and future directions for the field of motivation psychology: A special issue in honor of Prof. Dr. Willy Lens. *Psychologica Belgica*, 56, 317–341. <http://dx.doi.org/10.5334/pb.354>
- Vansteenkiste, M., Sierens, E., Soenens, B., Luyckx, K., & Lens, W. (2009). Motivational profiles from a self-determination perspective: The quality of motivation matters. *Journal of Educational Psychology*, 101, 671–688. <http://dx.doi.org/10.1037/a0015083>
- Vedder-Weiss, D., & Fortus, D. (2011). Adolescents' declining motivation to learn science: Inevitable or not? *Journal of Research in Science Teaching*, 48, 199–216. <http://dx.doi.org/10.1002/tea.20398>
- Vijendra, S., & Shivani, P. (2014). *Robust outlier detection technique in data mining: A univariate approach* (Report No. MT CA 2011). Retrieved from Cornell University Library website: <http://arxiv.org/ftp/arxiv/papers/1406/1406.5074.pdf>
- Virginia Department of Education (VDOE). (2012). *Title I, Part A: Improving basic programs operated by local educational agencies*. Retrieved from [http://www.doe.virginia.gov/federal\\_programs/esea/title1/part\\_a/](http://www.doe.virginia.gov/federal_programs/esea/title1/part_a/)
- Virginia Department of Education, Office of School Nutrition Programs. (VDOE SNP). (2014). *School year 2013–2014 National School Lunch Program (NSLP) free and reduced price eligibility report: School level*. Retrieved from [http://www.pen.k12.va.us/support/nutrition/statistics/free\\_reduced\\_eligibility/2013-2014/schools/frpe\\_sch\\_report\\_sy2013-14.pdf](http://www.pen.k12.va.us/support/nutrition/statistics/free_reduced_eligibility/2013-2014/schools/frpe_sch_report_sy2013-14.pdf)
- Voelkl, K. E. (1997). Identification with school. *American Journal of Education*, 105, 294–318. <http://dx.doi.org/10.1086/444158>
- von Eye, A., & Mun, E. Y. (2005). *Analysis of rater agreement: Manifest variable methods*. Mahwah, NJ: Psychology Press.
- Wang, M.-T., & Eccles, J. S. (2013). School context, achievement motivation, and academic engagement: A longitudinal study of school engagement using a multidimensional perspective. *Learning and Instruction*, 28, 12–23. <http://dx.doi.org/10.1016/j.learninstruc.2013.04.002>
- Weissman, J., & Magill, B. A. (2008). Developing a student typology to examine the effectiveness of first-year seminars. *Journal of the First-Year Experience & Students in Transition*, 20, 65–90.
- Wentzel, K. R. (1997). Student motivation in middle school: The role of perceived pedagogical caring. *Journal of Educational Psychology*, 89, 411–419. <http://dx.doi.org/10.1037/0022-0663.89.3.411>
- Wentzel, K. R., Battle, A., Russell, S. L., & Looney, L. B. (2010). Social supports from teachers and peers as predictors of academic and social motivation. *Contemporary Educational Psychology*, 35, 193–202. <http://dx.doi.org/10.1016/j.cedpsych.2010.03.002>
- Wentzel, K. R., & Wigfield, A. (2009). Introduction. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 1–8). New York, NY: Routledge.
- White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, 66, 297–333. <http://dx.doi.org/10.1037/h0040934>
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68–81. <http://dx.doi.org/10.1006/ceps.1999.1015>
- Wigfield, A., Eccles, J. S., Schiefele, U., Roeser, R. W., & Davis-Kean, P.



(2006). Development of achievement motivation. In N. Eisenberg (Ed.), *Handbook of child psychology* (Vol. 3, pp. 933–1002). New York, NY: Wiley.

*American Educational Research Journal*, 49, 124–154. <http://dx.doi.org/10.3102/0002831211426200>

Wormington, S. V., Corpus, J. H., & Anderson, K. G. (2012). A person-centered investigation of academic motivation and its correlates in high school. *Learning and Individual Differences*, 22, 429–438. <http://dx.doi.org/10.1016/j.lindif.2012.03.004>

Xu, J., Coats, L. T., & Davidson, M. L. (2012). Promoting student interest in science: The perspectives of exemplary African American teachers.

Received April 27, 2015

Revision received October 28, 2016

Accepted November 17, 2016 ■

### New Editors Appointed, 2019–2024

The Publications and Communications Board of the American Psychological Association announces the appointment of 11 new editors for 6-year terms beginning in 2019. As of January 1, 2018, new manuscripts should be directed as follows:

- *Clinician's Research Digest: Adult Populations and Clinician's Research Digest: Child and Adolescent Populations* (<http://www.apa.org/pubs/journals/crd/> and <http://www.apa.org/pubs/journals/crp/>), **Marisol Perez, PhD**, Arizona State University
- *Journal of Experimental Psychology: Learning, Memory, and Cognition* (<http://www.apa.org/pubs/journals/xlm/>), **Aaron S. Benjamin, PhD**, University of Illinois at Urbana-Champaign
- *Journal of Neuroscience, Psychology, and Economics* (<http://www.apa.org/pubs/journals/npe/>), **Samuel M. McClure, PhD**, Arizona State University
- *Journal of Threat Assessment and Management* (<http://www.apa.org/pubs/journals/tam/>), **Laura S. Guy, PhD**, Protect International Risk and Safety Services Inc., Vancouver, BC, Canada
- *Professional Psychology: Research and Practice* (<http://www.apa.org/pubs/journals/pro/>), **Kathi A. Borden, PhD**, Antioch University New England
- *Psychiatric Rehabilitation Journal* (<http://www.apa.org/pubs/journals/prj/>), **Sandra G. Resnick, PhD**, VA Connecticut Healthcare System
- *Psychology and Aging* (<http://www.apa.org/pubs/journals/pag/>), **Elizabeth A. L. Stine-Morrow, PhD**, University of Illinois at Urbana-Champaign
- *Psychology of Violence* (<http://www.apa.org/pubs/journals/vio/>), **Antonia Abbey, PhD**, Wayne State University
- *Psychology, Public Policy, and Law* (<http://www.apa.org/pubs/journals/law/>), **Michael E. Lamb, PhD**, University of Cambridge
- *Training and Education in Professional Psychology* (<http://www.apa.org/pubs/journals/tep/>), **Debora J. Bell, PhD**, University of Missouri-Columbia
- *Traumatology* (<http://www.apa.org/pubs/journals/trm/>), **Regardt J. Ferreira, PhD**, Tulane University

Current editors Thomas E. Joiner, PhD, Robert Greene, PhD, Daniel Houser, PhD, and Bernd Weber, PhD, Stephen D. Hart, PhD, Ronald T. Brown, PhD, Judith A. Cook, PhD, and Kim T. Mueser, PhD, Ulrich Mayr, PhD, Sherry Hamby, PhD, Michael E. Lamb, PhD, Michael C. Roberts, PhD, and Brian E. Bride, PhD, will receive and consider new manuscripts through December 31, 2017.

# Ethnic Composition and Heterogeneity in the Classroom: Their Measurement and Relationship With Student Outcomes

Camilla Rjosk

Humboldt-Universität zu Berlin, Germany

Dirk Richter

Humboldt-Universität zu Berlin, and University of Potsdam, Germany

Oliver Lüdtke

Leibniz Institute for Science and Mathematics Education at the University of Kiel (IPN), and Centre for International Student Assessment (ZIB), Germany

Jacquelynne Sue Eccles

University of California, Irvine

This study explores various measures of the ethnic makeup in a classroom and their relationship with student outcomes. We examine whether measures of ethnic diversity are related to achievement (mathematics, reading) and feeling of belonging with one's peers over and above commonly investigated composition characteristics. Multilevel analyses were based on data from a representative sample of 18,762 elementary school students in 903 classrooms. The proportion of minority students and diversity measures showed negative associations with student outcomes in separate models. Including diversity measures and the proportion of minority students, diversity of minority students mostly lost its significance. However, the results suggest that diversity measures may provide additional information over and above other classroom characteristics for some student outcomes. The various measures of diversity led to comparable results.

## *Educational Impact and Implications Statement*

This study suggests that the ethnic makeup of classrooms is related to student outcomes. That is, students in classes with a higher proportion of ethnic minority students showed slightly lower achievement and feeling of belonging with one's peers even if the socioeconomic status, the immigrant background of the family, cognitive ability, and gender of the student is equal. In addition to the proportion of ethnic minority students, average socioeconomic status, and average cognitive abilities, we looked at the ethnic heterogeneity in each classroom and found that this was mostly independent from student outcomes. Only for math we found a positive association indicating that students in a more ethnically diverse classroom showed slightly higher test scores—however, this slight association cannot be interpreted as a causal relationship because of our cross-sectional design. The findings suggest that measures of heterogeneity may uncover relationships that the mere proportion of minority students which disregards various ethnic groups in the classroom is unable to show and open a discussion on how to investigate effects of ethnic diversity in educational research.

**Keywords:** ethnic composition, diversity, multilevel analysis, academic achievement, classroom

**Supplemental materials:** <http://dx.doi.org/10.1037/edu0000185.supp>

As societies become more diverse in terms of ethnic background, the composition of the student body within educational systems diversifies as well. This ethnic makeup of schools and

classrooms can be described by two different characteristics that are associated with each other: the proportion of minority students and ethnic heterogeneity or diversity. Both characteristics have

This article was published Online First March 23, 2017.

Camilla Rjosk, Institute for Educational Quality Improvement (IQB), Humboldt-Universität zu Berlin; Dirk Richter, Institute for Educational Quality Improvement (IQB), Humboldt-Universität zu Berlin, and Institute for Educational Sciences, University of Potsdam; Oliver Lüdtke, Department of Educational Measurement, Leibniz Institute for Science and Mathematics Education at the University of Kiel (IPN), and Centre for International Student Assessment (ZIB), Germany; Jacquelynne Sue Eccles, School of Education, University of California, Irvine.

During the work on her dissertation, Camilla Rjosk was a predoctoral fellow of the International Max Planck Research School on the Life Course (LIFE, [www.imprs-life.mpg.de](http://www.imprs-life.mpg.de); participating institutions: MPI for Human Development, Freie Universität Berlin, Humboldt-Universität zu Berlin, University of Michigan, University of Virginia, University of Zurich).

Correspondence concerning this article should be addressed to Camilla Rjosk, Institute for Educational Quality Improvement (IQB), Humboldt-Universität zu Berlin, Germany. E-mail: [Camilla.Rjosk@iqb.hu-berlin.de](mailto:Camilla.Rjosk@iqb.hu-berlin.de)



been used in research to describe school and classroom settings and to investigate the relationships with student achievement. In particular, research focusing on associations with student outcomes assumes that the composition of the student body shapes the learning environment and therefore also the outcome of student learning.

Various theoretical accounts assume a negative relationship between the proportion of ethnic minority students and student achievement based on school resources and mediated by instructional quality, language spoken with peers, and learning culture (Driessen, 2002; Goldsmith, 2011; Raudenbush, Fotiu, & Cheong, 1998; Stipek, 2004). In addition, several authors assume positive effects of ethnic heterogeneity on achievement because students in heterogeneous learning environments encounter and have to work through contradictions and discrepancies in everyday life and therefore may be able to expand their intellectual capacities (e.g., Benner & Crosnoe, 2011; Gurin, Dey, Gurin, & Hurtado, 2003; Peetsma, Van der Veen, Koopman, & Van Schooten, 2006; Tam & Bassett, 2004).

Research exploring the relationship between the ethnic makeup of schools or classrooms and student achievement shows mixed results: The proportion of ethnic minority students in a school or classroom often has no or slightly negative predictive effects on student achievement (Mickelson, Bottia, & Lambert, 2013; Van Ewijk & Sleegers, 2010a). For ethnic heterogeneity, some studies report that a higher proportion of ethnically heterogeneous students may lead to higher achievement (e.g., Benner & Crosnoe, 2011; Tam & Bassett, 2004).

Most studies have dealt only with a broad distinction between ethnic minority and majority without addressing and measuring ethnic heterogeneity or diversity, yet their authors sometimes interpret the results in the light of diversity. The present study compares various measures of ethnic composition and heterogeneity used in different disciplines with the goal of better understanding the relationship between the ethnic makeup of classrooms and student outcomes. Our aim is to investigate whether the measures of ethnic diversity are related to student achievement and psychosocial outcomes over and above commonly investigated characteristics of classroom composition. That is, we want to find out whether the proportion of minority students is sufficient to describe effects of the ethnic makeup or whether diversity measures can provide additional information.

### Relationship Between Characteristics of the Student Body and Individual Student Outcomes

Students differ in their educational success and level of achievement outcomes. This variability is associated with individual background characteristics, such as cognitive abilities, prior knowledge, and the socioeconomic background of their families and associated home learning environment (e.g., OECD, 2010). In addition to these individual and family characteristics, the composition of the student body matters for individual outcomes. Although some authors state that compositional effects may reflect mere methodological artifacts (e.g., Hauser, 1970), current research concludes, for instance, that students tend to show higher achievement in classrooms that are characterized by a high average prior achievement level and a high average socioeconomic status (SES) of the student body (Van Ewijk & Sleegers, 2010b). How-

ever, research is inconclusive on whether the ethnic composition is related to student outcomes (e.g., Driessen, 2002), independent of the average prior achievement and average SES of the classroom.

The present article focuses on the relationship between the ethnic makeup of classrooms and students' achievement, as well as psychosocial outcomes. More precisely, the term ethnic makeup may pertain to two characteristics that represent different strands of theory and research: (a) the proportion of ethnic minority students as the measure of ethnic composition commonly used in educational research and (b) ethnic heterogeneity measured by various indices analogous to the concept of diversity operationalized in a large number of different disciplines.

### Definitions: Ethnic Composition and Heterogeneity

Educational research that addresses questions of the ethnic makeup of classrooms commonly operationalizes the ethnic composition by calculating the *proportion of ethnic minority students* in a classroom or school. For instance, international meta-analyses on ethnic composition and student achievement with about 38 primary studies consistently distinguish between ethnic minority and majority students or single minority groups and majority students (Mickelson et al., 2013; Van Ewijk & Sleegers, 2010a). A compositional effect here reflects the effect of the proportion of minority students even after controlling for the effect of the individual minority background (see Raudenbush & Bryk, 2002). This approach to diversity is sometimes referred to as "simplistic majority-minority approach" (Budescu & Budescu, 2012), because it only draws a superficial picture of the actual ethnic classroom composition. Related to the ethnic composition is the idea of heterogeneity: It is often implicitly assumed that a high proportion of ethnic minority students represents a heterogeneous student body. However, the majority-minority distinction does not provide much information on heterogeneity because it disregards the distribution of various ethnic groups.

The concept of *heterogeneity* or *diversity* plays a key role in a large number of disciplines, such as ecology (e.g., McCann, 2000), economics (e.g., Hall & Tideman, 1967), organizational psychology (e.g., Hoppe, Fujishiro, & Heaney, 2014; Meyer, in press), communication (e.g., Dimmick & McDonald, 2001), and geography (e.g., Les & Maher, 1998). Their operationalizations of diversity can be used in educational research as well. Yet, they are less common in this field (as an example, see Benner & Crosnoe, 2011). In the current study, diversity is understood as "the distribution of population elements along a continuum of homogeneity to heterogeneity with respect to one or more variables" (Lieberson, 1969, p. 851 cited in Budescu & Budescu, 2012; cf. Teachman, 1980). This concept is also referred to as "variety diversity" describing differences in group compositions according to a categorical variable (see Harrison & Klein, 2007, for a classification of diversity concepts). In contrast to compositional effects mentioned previously, diversity is a mere classroom level characteristic. Existing operationalizations capture the information of diversity with a measure that represents a single, dual, or threefold concept (see Junge, 1994; McDonald & Dimmick, 2003; Stirling, 2007). That is, the measure includes one or more of the following pieces of information: number of categories, distribution of elements across categories, and a numerical distance measure that expresses how similar the various categories are to each other. These pieces of information are combined using relative frequencies (e.g., Simpson's



*D*) or logarithms of those frequencies (e.g., Shannon's *H*). The present study applies single and dual concept measures that address only the number of categories (i.e., number of ethnicities present in a classroom) or additionally the distribution of elements across categories (i.e., how many students of each ethnicity are there in a classroom, for instance included in the measures Simpson's *D* and Shannon's *H*).<sup>1</sup> These diversity indices thus provide more information on heterogeneity than the majority–minority distinction because they reflect the multitude of various ethnic backgrounds. A detailed overview of the most common diversity measures can be found in Table S.1, available as online supplemental material.<sup>2</sup>

These measures and their comparison have received relatively little attention in educational research thus far. Reviews in other fields showed that some measures are more sensitive to the number of categories (e.g., Junge's *H*) or to changes in the largest proportion of categories (e.g., Simpson's *D*) than other measures, and that the measures are close in agreement when using them to quantify diversity (McDonald & Dimmick, 2003). McDonald and Dimmick (2003) concluded that the measures Simpson's *D* and Shannon's *H* are most appropriate if one is interested in a measure that is simultaneously sensitive to the number of categories and the maximum proportion of categories. A recent review by Budescu and Budescu (2012) in the educational field that explored two diversity measures (Simpson's *D* and Shannon's *H*) found that ranking schools according to these two indices did not lead to differences in ranking. However, Shannon's *H* was related slightly stronger to school-level achievement than Simpson's *D*. Overall, there is a dearth of research examining the applicability of different diversity measures in the educational field and how they are related to student outcomes. Therefore, the present study starts by describing a variety of different diversity measures within the preliminary analyses section (see "Comparison and selection of diversity measures") before including these measures into the analysis models.

### Proportion of Ethnic Minority Students and Individual Student Outcomes

Theories and empirical research on the proportion of ethnic minority students typically address effects on student achievement and assume negative relationships between these two characteristics. Research does not suggest that the proportion of minority students per se is related to student outcomes, rather that there are mediating processes and associated aspects that induce these relationships. Interrelated factors that can explain why the proportion of ethnic minority students— independent of the prior achievement and socioeconomic composition as well as individual background characteristics—may be negatively related to student achievement outcomes are as follows: (a) school resources, (b) instructional quality, (c) minority language usage, and (d) learning culture.

First, ethnic minority students often have less access to schools with good resources and favorable organizational and structural features such as class size, teacher qualifications, and programs that encourage learning. They are more likely to attend residential neighborhood schools with poor resources in the segregated areas they live in (Raudenbush et al., 1998).

Second, classes with high proportions of ethnic minority students may encounter less beneficial learning opportunities in terms of instructional quality; for example, less challenging tasks and a less student-oriented climate. This is based on the assumption that

teachers show lower achievement expectations toward ethnic minority students (Ready & Wright, 2011), which, in turn, may cause them to offer fewer challenging learning opportunities and to engage in less positive interactions with these students (for a meta-analysis and review, see Den Brok & Levy, 2005; Tenenbaum & Ruck, 2007). In addition, teachers in segregated neighborhood schools with poor resources that ethnic minority students frequent tend to be less qualified to deliver high-quality instruction. These inequalities arise, for example, because the most qualified teachers gradually shift to less-disadvantaged schools within an area because they typically have first right of transfer when vacancies appear (Betts, Rueben, & Danenberg, 2000).

Third, classrooms with high proportions of ethnic minority students also tend to have high proportions of students who do not speak the language of instruction at home. As a consequence, these students may be less able to support each other by explaining learning materials in the language of instruction. Furthermore, they may not speak the language of instruction with each other in situations such as school breaks, which results in fewer learning opportunities and may negatively affect students' language-related achievement (see Driessen, 2002; Entwisle & Alexander, 1994; Peetsma et al., 2006; Van Ewijk & Sleegers, 2010a).

Fourth, ethnic minority students may share values, beliefs, and behaviors associated less with learning and achievement (e.g., negative attitudes toward school, pessimism, and irregular school attendance). Originally focusing on motivation of Afro-American students in the United States, Ogbu's (1987) cultural ecological theory assumes that minority students are assigned a subordinate status, devaluated in school and feel vulnerable. As a consequence, these students may come to feel alienated from school, to reject educational values, and to be less motivated to learn (see Kumar & Maehr, 2010; Ogbu, 2004). In classes with a large number of such students, peers transmit these values and beliefs through interacting with each other. Thus, a less beneficial learning culture negatively affecting motivation to learn and achievement may emerge (see Agirdag, Van Houtte, & Van Avermaet, 2012; Goldsmith, 2011).

The aforementioned factors would predict a negative relationship between the proportion of minority students and student achievement, but a few theories also assume a positive relationship with some psychosocial outcomes, such as feelings of belonging with one's classmates and learning motivation. According to the self-determination theory, social relatedness or belongingness is one of the three basic needs whose fulfillment is assumed to foster motivation to learn (Deci & Ryan, 2000; Niemiec & Ryan, 2009). Motivation to learn and feeling of belonging with one's classmates are relevant educational goals of schooling in addition to achievement development because they are associated with a number of favorable outcomes, such as engagement in learning activities, higher school achievement levels, keeping track in school, life satisfaction, and mental health. Especially for minority students,

<sup>1</sup> The present study does not include threefold concepts of diversity measures because we do not focus on numerical distance measures expressing similarity or dissimilarity between countries of birth or ethnicities and because there are no appropriate distance measures available.

<sup>2</sup> The present study covers diversity as a characteristic at the classroom level and does not include operationalizations such as the proportion of students with the same background as an individual student at the student level (see Benner & Crosnoe, 2011; Hoppe et al., 2014).



minority versus majority group membership may act as a lens through which individuals in a culturally pluralistic society view each other and on which they build their sense of belonging (see belongingness perspective by Baumeister & Leary, 1995; Johnson, Crosnoe, & Elder, 2001; Kumar & Maehr, 2010). This assumption mostly refers to students with the same specific ethnic background as a source of belongingness, yet it may also apply to majority versus minority group membership in society. According to social identity theory (Tajfel & Turner, 1986) and the similarity–attraction paradigm (Byrne, 1971), group membership and feeling of belonging are based on similarity between students. Minority students may view themselves to be more similar to each other than to majority student; for instance, in terms of multilingual experiences and immigration history within the family. In their review, Kumar and Maehr (2010) argue that minority students often feel rejected by peers from the majority group in society and therefore show less motivation to learn. By implication, a classroom with a high proportion of minority students should strengthen minority students' feeling of belonging with their peers and facilitate their motivation to learn. This could result in a u-shaped relationship between the proportion of minority students and students' feeling of belonging at the classroom level. In addition, a heightened feeling of belonging with one's classmates could positively affect future student achievement (Christenson, Reschly, & Wylie, 2012). Some authors also assume a linear positive relationship between the proportion of minority students and motivational characteristics. These assumptions include, for instance, that classrooms with low proportions of minority students may be more competitive and focus more on performance goals that are less favorable for achievement development. They also focus on specific immigrant groups, such as Asian American students in the United States, who may share especially heightened learning motivation (cf. Zusho, Pintrich, & Cortina, 2005).

International meta-analyses (Mickelson et al., 2013; Van Ewijk & Sleegers, 2010a) commonly find a substantial but small negative effect of the proportion of minority students in predicting student achievement. This effect varies in size depending on the minority groups explored, the students' age, the control variables included, and the operationalization of the constructs. Although more research on mediating processes is clearly needed, some findings support the hypotheses of ethnic inequalities in access to schools with high resources (Raudenbush et al., 1998), of lower instructional quality in classrooms with high proportions of ethnic minority students (Palardy, 2015; Stipek, 2004), and a less favorable learning culture in schools with high proportions of ethnic minority students and average low SES (Agirdag et al., 2012; Goldsmith, 2011). In addition, a few studies suggest that minority students look forward to instruction more and believe that the topics they learn will be more useful in the future if they attend schools with high proportions of minority students (Goldsmith, 2004). Furthermore, language minority students showed to be more motivated to learn in language lessons in classrooms with higher proportions of language minority students (Rjosk, Richter, Hochweber, Lüdtke, & Stanat, 2015).

### **Ethnic Diversity and Individual Student Outcomes**

According to Piaget's (1977) concept of disequilibrium, ethnic diversity should have positive effects on students' cognitive development. Being faced with new information that does not fit into one's

schemas—for instance, through exposure to multiple perspectives from people with varying ethnic backgrounds—induces a state of unpleasantness. This state drives the learning process through assimilation and accommodation of new ideas and thus fosters cognitive development. As a consequence, several authors assume positive effects of ethnic diversity in the classroom on students' achievement (Benner & Crosnoe, 2011; Gurin et al., 2003; Tam & Bassett, 2004). This line of argument is similar to the information/decision-making perspective taken in organizational psychology to explain positive effects of work team diversity (see Meyer, *in press*). Studies in educational research found, for instance, that students in ethnically more heterogeneous kindergartens showed higher achievement levels in mathematics and reading, after controlling for the socioeconomic composition, proportion of minority students, and other school characteristics (Benner & Crosnoe, 2011). Students from ethnically diverse high schools also showed university grade point averages (GPAs) in the first semester that were one-fourth to one-half point higher than that of students from a nondiverse high school, after controlling for achievement composition and quality of high schools (Tam & Bassett, 2004).

Although positive relationships are assumed between ethnic diversity on the one hand and cognitive development as well as school achievement on the other, a negative association between diversity and students' feeling of belonging with their classmates may exist (Benner & Crosnoe, 2011; Benner, Graham, & Mistry, 2008). According to the belongingness perspective mentioned previously (see Baumeister & Leary, 1995; Byrne, 1971; Tajfel & Turner, 1986), ethnic diversity should be negatively associated with attachment to the peers. These theories assume that students' sense of belonging is more strongly related to the specific ethnicity of the peers than to the broader category of minority status. For instance, in a study on U.S. ninth-graders, students perceived the school climate to be fairer and more directed toward academics and interracial understanding if they attended ethnically less diverse schools (Benner et al., 2008). Furthermore, ethnically less diverse schools and classrooms were characterized by stronger attachment to school (Johnson et al., 2001) and lower levels of perceived cultural discrimination (Seaton & Yip, 2009).

In sum, educational research commonly assumes a negative relationship between the proportion of ethnic minority students and student achievement. Empirical findings indicate that the relationships may be as predicted, yet there is not much research on the underlying mechanisms inducing compositional effects. However, some studies also find a positive relationship between ethnic classroom diversity and students' achievement. For students' feeling of belonging with their classmates as a psychosocial outcome, theories predict a negative relationship between heterogeneity and feeling of belonging and a u-shaped relationship between the proportion of minority students and feeling of belonging. While the former assumption is supported by some studies, the latter has not been investigated to our knowledge.

To conclude, the two strands of theories and findings presented so far would lead to contradictory predictions: A high proportion of ethnic minority students in a classroom has been shown to be negatively related to student achievement. Simultaneously, a high proportion of ethnic minority students partly corresponds to a higher ethnic diversity which is assumed to be positively related to school achievement. This raises the question of how the ethnic composition is related to school achieve-



ment when one tries to disentangle the effects of the proportion of ethnic minority students and ethnic diversity. Furthermore, we are interested in the question of how ethnic diversity and the broader distinction between ethnic majority and ethnic minority students are related to students' feeling of belonging with their classmates as an example of a non-cognitive student outcome. Because theories suggest that the feeling of belonging to one's classmates should be positively related to student achievement (Christenson et al., 2012), and could possibly mediate the relationship between classroom composition and achievement outcomes, we will further explore these relationships in a full model including various student outcomes.

### The Present Study: Research Questions and Hypotheses

The aim of the present study is to examine the relationship between the ethnic makeup of classrooms and student achievement in mathematics and reading comprehension and students' feeling of belonging with their classmates. The competing theoretical assumptions and empirical findings presented in the last sections form the basis of our study. We examine the relationships between student outcomes and the proportion of minority students, as well as ethnic diversity in German elementary school classrooms in a cross-sectional design. That is, our analyses provide information on associations between classroom characteristics and student outcomes, but do not allow drawing conclusions about causal effects. The average SES in a classroom and the average prior achievement are classroom-level background variables that have been shown to matter for individual achievement outcomes in former studies (e.g., Van Ewijk & Sleegers, 2010b). Our analyses include these variables as covariates. We take the average prior achievement into account using the average cognitive abilities in the classroom as a proxy. Our research questions and hypotheses are as follows:

**(1) Relationship with achievement scores.** **(1a) Is the proportion of ethnic minority students in a classroom related to students' achievement in mathematics and reading comprehension?** We predict that the proportion of minority students will be negatively related to achievement outcomes. The underlying assumption is that in classrooms with high proportions of minority students, there is a less favorable learning environment characterized by poor school resources, lower instructional quality, non-German language usage with peers, and less favorable learning culture (see "Proportion of ethnic minority students and individual student outcomes").

**(1b) Is ethnic diversity in the classroom—operationalized by various diversity measures—related to students' achievement in mathematics and reading comprehension?** Contrary to some studies reviewed in the last sections, we predict a negative relationship between ethnic diversity measures and achievement for the same reasons as those as described in Hypothesis 1a.

**(1c) Do diversity measures explain additional variance in student achievement over and above the proportion of ethnic minority students?** That is, we ask whether the proportion of minority students is sufficient to investigate associations with the ethnic makeup in terms of diversity or whether additional measures provide further information. We predict that measures of ethnic diversity provide additional information on the classroom composition and therefore will be related to student achievement

outcomes over and above the proportion of ethnic minority students in a classroom. We furthermore assume that controlling for the proportion of ethnic minority students in a classroom and other background characteristics also controls for characteristics of the learning environment associated with classroom composition. Holding that constant might offer the opportunity to investigate whether diversity is positively associated with achievement as predicted in the literature. We predict additional positive relationships between achievement scores and ethnic diversity.

**(2) Relationship with feeling of belonging with one's peers.** **(2a) Is the proportion of ethnic minority students related to students' feeling of belonging with their classmates?** We assume that the proportion of ethnic minority students reveals different relationships with the feeling of belonging for minority students than for majority students. According to the belongingness hypothesis, minority students should feel more attached in classrooms with high proportions of minority students, and majority students should feel more attached in classrooms with high proportions of majority students. That is, we predict an interaction between the proportion of minority students and individual minority status. In line with this assumption, the proportion of minority students should be on average related to the feeling of belonging with one's peers in a u-shaped manner.

**(2b) Is ethnic diversity in the classroom—operationalized by various diversity measures—related to students' feeling of belonging with their classmates?** We assume that ethnic diversity is negatively related to the feeling of belonging because a large diversity in a classroom corresponds to a low number of students from the same ethnic background (see "Ethnic diversity and individual student outcomes").

**(2c) Do diversity measures explain additional variance in students' feeling of belonging with the peers over and above the proportion of ethnic minority students?** That is, we ask whether the proportion of minority students is sufficient to investigate associations with the ethnic makeup in terms of diversity or whether additional measures provide further information. We assume that ethnic diversity explains additional variance in students' feeling of belonging with the peers over and above the proportion of ethnic minority students as students should build their sense of belonging on the specific ethnicity of the peers in their classroom rather than on the broader category of minority status (see "Ethnic diversity and individual student outcomes").

As a last step, we explore the relationship between the ethnic makeup of classrooms, achievement measures, and students' feeling of belonging with their classmates in a full model (i.e., including both ethnic makeup and feeling of belonging as predictors of achievement) to gain first insights into possibly mediating effects of students feeling of belonging.

## Method

### Participants

Our analyses are based on data from a nationally representative sample of elementary school students in Germany who participated in the 2011 National Assessment Study of student achievement in elementary schools (*IQB-Ländervergleich*; Stanat, Pant, Böhme, & Richter, 2012) of the German Institute for Educational Quality Improvement (IQB). The data include 27,081 students of complete



fourth-grade classrooms in 1,349 randomly selected German public schools (see Richter et al., 2012).

We excluded special needs schools, schools from the former German Democratic Republic (GDR)—because they have very low proportions of ethnic minority students (see Federal Statistical Office Germany, 2012)—and classrooms with a large number of missing values for ethnic background information (see “Missing data treatment”). As a consequence, our analyses were based on 18,762 students attending 903 classrooms in 903 schools (average number of students per classroom,  $M = 21$ ). For a sample description, see Table 1.

## Measures

We use information from standardized achievement tests, student questionnaires, and parent questionnaires, which were completed anonymously.

### Student-level independent variables.

**Ethnic background of students.** We categorized the ethnic background of students using information from the parent questionnaire. If the parent response was missing, we used information from the student questionnaire (see “Missing data treatment”). The ethnic background was categorized based on the country where the parents were born. To be categorized as a student with minority status, at least one parent had to be born abroad. For instance, a student with a Turkish background has parents who were both born in Turkey or one parent in Turkey and one parent in Germany. If one parent was born in Turkey and one parent in another country (i.e., not Germany), the student was assigned to the category “other country.” Within the category “other country,” countries most represented were Iran and Arab countries. In our analyses, we distinguish six groups (Table 1) corresponding to the largest groups in this sample (see Stanat et al., 2012). If questionnaire information was completely missing, we did not exclude the student—because it would distort the classroom-level analyses—but assigned him or her to the category “unidentifiable,” which we included in the analyses (Table 1). For analyses comparing broadly ethnic majority and ethnic minority students, students with a German background (i.e., both parents born in Germany) were categorized as “ethnic majority” and the remaining groups as “ethnic minority.”

**Students’ SES.** As the measure of SES we used the highest International Socio-Economic Index of Occupational Status (HISEI; Ganzeboom, 2010; Ganzeboom, De Graaf, Treiman, & De Leuw, 1992). This index is a classification of parents’ occupation based on income and education, with a score range of 10 (e.g., a kitchen helper) to 89 (e.g., a medical doctor). We used information about the current occupation that the parents provided in the questionnaire in an open answer format.

**Cognitive abilities.** To approximate students’ prior achievement within the cross-sectional design, we used the figural subtest of the cognitive abilities test (KFT 4–12 + R; Heller & Perleth, 2000; see Baumert, Stanat, & Watermann, 2006, on the validity of such a proxy). This standardized test is a commonly used cognitive abilities test in Germany and comparable to the cognitive abilities test (CAT) by Thorndike and Hagen (1971, 1993). We use the figural analogies subscale consisting of 25 items in which students are asked to choose one figure out of five possibilities in analogy to a given pair of figures. The test authors describe the reliability and validity of this measure as very satisfactory (internal consistency of figural subscale:  $\alpha = .92$ ; retest reliability of the whole test consisting of figural, verbal, and numerical parts after 2 years:  $r_{tt} = .83$ ; average predictive validity of the whole test: correlation of  $r = .41$  with GPA of higher education entrance certification up to 8 years later). Internal consistency of the subscale in the analysis sample was  $\alpha = .93$ . Correlations with achievement outcome measures in the analysis sample were  $r_{\text{mathematics}} = .54$  and  $r_{\text{reading}} = .49$ .

**Gender.** Student gender was recorded in the tracking form completed by the classroom teacher (dummy coding; male = 0, female = 1).

**Classroom level independent variables.** The key classroom level variables of this study were the proportion of ethnic minority students in a classroom and ethnic diversity. Covariates were the classroom level SES and cognitive abilities.

**Proportion of ethnic minority students.** We calculated the proportion of ethnic minority students in each classroom in accordance with the individual student level categorizations described

Table 1  
*Descriptive Sample Statistics for Demographic Variables*

Variable	<i>M (SD)</i>	Minimum	Maximum
Individual level (L1), <i>N</i> = 18,762			
Ethnic minority status	37.60%	—	—
SES	50.23 (16.19)	10	89
Female	49.50%	—	—
Age in years	10.41 (.50)	6.83	13.17
Classroom level (L2), <i>N</i> = 903			
Proportion of German background	60.52% (22.08)	0%	100%
Proportion of Turkish background	6.53% (9.50)	0%	61.11%
Proportion of former USSR background	4.91% (7.63)	0%	55.56%
Proportion of Polish background	2.13% (4.00)	0%	41.67%
Proportion of former Yugoslavia background	2.96% (4.61)	0%	29.41%
Proportion of other background	9.71% (9.09)	0%	52.94%
Proportion of students with missing background information (“unidentifiable”)	13.24% (10.67)	0%	48.15%
Proportion of ethnic minority students	39.48% (22.08)	0%	100%
SES	49.64 (8.15)	25.75	78.07

*Note.* SES = socioeconomic status. For operationalization of SES and students’ ethnic background, see Measures section.

previously as relative frequency (see the next paragraph “Ethnic diversity” for an example).

**Ethnic diversity.** We operationalized ethnic classroom diversity calculating various diversity measures (Table S.1) based on individual ethnic background information described previously. In the following, we explain how four exemplary measures to operationalize the ethnic makeup of classrooms are computed: the proportion of ethnic minority students and three diversity measures—number of categories, Simpson’s *D*, and Shannon’s *H*.

Imagine two fictitious classrooms, Classroom A and Classroom B. Both classrooms have 20 students. In Classroom A, there are 6 students with German background, 13 with Turkish background, and 1 with Polish background. In Classroom B, there are 6 students with German background, 5 with Turkish background, 4 with parents from the former USSR, 4 with parents from the former Yugoslavia, and 1 with another ethnic background. The *proportion of ethnic minority students* in Classroom A is 0.7 ( $\text{prop}_A = (13 + 1)/20 = 0.7$ ), and in Classroom B it is also 0.7 ( $\text{prop}_B = (5 + 4 + 4 + 1)/20 = 0.7$ ). The *number of ethnic groups* in Classroom A equals 3 ( $N_{cat_A} = \text{German, Turkish, Polish} = 3$ ), and in Classroom B it equals 5 ( $N_{cat_B} = \text{German, Turkish, former USSR, former Yugoslavia, other} = 5$ ). We now use this scenario to illustrate further diversity measures.

*Simpson’s D* represents the probability that two students selected at random from a classroom belong to different ethnicities—thus the greater the value of *D*, the greater the diversity. Its minimum value is 0, and its maximum is achieved when the distribution across the *c* ethnic groups in the classroom is uniform (in our study,  $c = 7$ , i.e.,  $D_{\text{max}} = (7 - 1)/7 = 0.857$ ). *Shannon’s H* involves a logarithmic transformation of probabilities ( $H_{\text{min}} = 0$ ;  $H_{\text{max}}$  in our study =  $\ln(c) = \ln(7) = 1.946$ ). *Simpson’s D* and *Shannon’s H* both use the relative frequencies ( $p_i$ ) of each ethnic group in the classroom. For Classroom A, the relative frequencies are  $p_{A,\text{German}} = 0.3$ ,  $p_{A,\text{Turkish}} = 0.65$ , and  $p_{A,\text{Polish}} = 0.05$ , and for Classroom B they are  $p_{B,\text{German}} = 0.3$ ,  $p_{B,\text{Turkish}} = 0.25$ ,  $p_{B,\text{USSR}} = 0.2$ ,  $p_{B,\text{Yugoslavia}} = 0.2$ , and  $p_{B,\text{other}} = 0.05$ . The measure *Simpson’s D* for Classroom A equals  $D_A = 0.485$  ( $D_A = 1 - \sum p_i^2 = 1 - ((0.3 \times 0.3) + (0.65 \times 0.65) + (0.05 \times 0.05)) = 1 - 0.515 = 0.485$ ) and for Classroom B  $D_B = 0.765$  ( $D_B = 1 - \sum p_i^2 = 1 - ((0.3 \times 0.3) + (0.25 \times 0.25) + (0.2 \times 0.2) + (0.2 \times 0.2) + (0.05 \times 0.05)) = 1 - 0.235 = 0.765$ ). The measure *Shannon’s H* for Classroom A equals  $H_A = 0.791$  ( $H_A = - \sum p_i \ln(p_i) = - ((0.3 \times \ln(0.3)) + (0.65 \times \ln(0.65)) + (0.05 \times \ln(0.05))) = 0.791$ ) and for Classroom B it equals  $H_B = 1.501$  ( $H_B = - \sum p_i \ln(p_i) = - ((0.3 \times \ln(0.3)) + (0.25 \times \ln(0.25)) + (0.2 \times \ln(0.2)) + (0.2 \times \ln(0.2)) + (0.05 \times \ln(0.05))) = 1.501$ ).

In sum, our example shows that, even though the proportion of minority students in Classroom A and B are the same, their student body differs in terms of ethnic homogeneity. Classroom B is ethnically more diverse than Classroom A, which is reflected in a larger number of ethnicities in the classroom and higher values of *Simpson’s D* and *Shannon’s H* (Table 2).

**Classroom level socioeconomic status.** We aggregated the average SES score for each classroom based on the individual student scores (see “Student-level independent variables”).

**Classroom level cognitive abilities.** We aggregated the average cognitive abilities test score for each classroom as a proxy for

Table 2  
*Ethnic Makeup of Two Different Classrooms:  
Exemplary Computations*

Variable	Classroom A	Classroom B
Proportion of minority students	.70	.70
Number of ethnic groups	3	5
Simpson’s <i>D</i>	.485	.765
Shannon’s <i>H</i>	.791	1.501

*Note.* Both classrooms involve 20 students but differ in their distribution of ethnic makeup (Classroom A: 6, 13, and 1; Classroom B: 6, 5, 4, 4, and 1).

prior achievement based on the individual student scores (see “Student-level independent variables”).

**Outcome variables.**

**Student achievement in reading comprehension and mathematics.** Trained test administrators conducted standardized achievement tests in German reading comprehension and mathematics in the classrooms. The tests were designed by a team of experienced teachers and scientists in partnership with the German Institute for Educational Quality Improvement (IQB) to measure these achievement domains in accordance with the German national educational standards. These educational standards were introduced by the *Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany* (Kultusministerkonferenz [KMK]). They describe competencies students are expected to have developed at a certain grade. The Institute for Educational Quality Improvement is responsible for coordinating the test development process, working with teachers and experts in subject-matter education, and evaluating the psychometric item properties based on pilot studies with nationally representative data sets.

Task units measuring reading comprehension consisted of a literary or factual text of half a page to one and a half pages and several items with varying complexity. The items were mainly presented as multiple-choice and short-answer questions.<sup>3</sup> Each student received two to four out of 11 task units (booklet design).

The mathematics achievement test covered the five content domains “numbers and operations,” “space and shape,” “patterns and structures,” “measurement,” and “probability” using a variety of tasks, such as simple computations, extracting information from charts, and reflecting shapes (see Winkelmann, van den Heuvel-Panhuizen, & Robitzsch, 2008).

A generalized Rasch model was used to estimate student achievement scores on a common scale for each achievement domain of reading and mathematics achievement. Mathematics and reading scores were generated using the plausible values (PV) technique (Adams, Wu, & Carstensen, 2007). Fifteen PVs were generated for each student for mathematics and reading achievement, which were scaled to have a mean score of 500 and an *SD* of 100 in the German student population (current sample distribution mathematics:  $M = 489.05$ ,  $SD = 96.00$ , reading:  $M = 494.12$ ,  $SD = 90.87$ ). Expected a posteriori reliabilities of plausible values (EAP/PV reliabilities; cf. Adams, 2005) in the calibration model were .91 (mathematics) and .73 (reading).

<sup>3</sup> For illustrative examples of test items, see [www.iqb.hu-berlin.de/laendervergleich/LV2011/Beispielaufgaben](http://www.iqb.hu-berlin.de/laendervergleich/LV2011/Beispielaufgaben).



**Feeling of belonging with one's peers.** Students rated their feeling of belonging with their classmates in the student questionnaire. The scale is part of a questionnaire measuring emotional and social experiences in school (Rauer & Schuck, 2003). It consists of four items rated on a 4-point Likert scale (1 = “fully disagree” to 4 = “fully agree”) by all students. Item examples are “My classmates are nice to me” and “When I am sad, my classmates comfort me” (manifest  $M = 3.3$ ,  $SD = 0.6$ , Cronbach's  $\alpha = .71$ ). We used these items to represent the latent construct of belonging with one's peers in a multilevel structural equation model framework (see “Data analysis” and “Measurement model fit and variance of outcome variables between classrooms”).

## Data Analysis

**Preliminary analyses.** As part of our preliminary analyses, we calculated eight diversity measures (Table S.1) and analyzed their relationship with each other using Pearson correlations (see “Comparison and selection of diversity measures”).

**Analyses for Research Questions 1 and 2.** We used structural equation modeling to explore our Research Questions 1 and 2 (see Bovaird, 2007). Several random intercept multilevel structural equation models with fixed slope were estimated to analyze the relationship between (a) the proportion of ethnic minority students in a classroom and student outcomes, (b) ethnic diversity operationalized by various measures and student outcomes, as well as (c) both classroom characteristics and student outcomes. We employed a stepwise model building procedure. We used the software Mplus (Version 6.1; Muthén & Muthén, 1998–2010) for all analyses with the option “type = imputation” pooling the results of analyses with the 15 plausible values of the student achievement measures.

Metric background variables at the student level (SES, cognitive abilities) were standardized, which implies centering at their grand mean. Categorical variables (ethnic minority status, gender) were neither centered nor standardized. The estimates of background variables aggregated at the classroom level, that is proportion of minority students, average SES, and average cognitive abilities, can be interpreted as compositional effects (Raudenbush & Bryk, 2002). A compositional effect reflects the effect of the aggregate of a person-level characteristic (e.g., proportion of minority students) even after controlling for the effect of the individual characteristic (e.g., individual minority background; see Raudenbush & Bryk, 2002). Effects of diversity indices do not show compositional effects but effects at the classroom level. All classroom level variables were standardized at the classroom level. We additionally included the quadratic term of the proportion of ethnic minority students in a classroom, in order to explore a potential nonlinear relationship between the proportion and student outcomes across classrooms. Prior to calculating the quadratic term, we centered the proportion of ethnic minority students at its mean to counteract multicollinearity. All regression coefficients in the result tables were standardized using the total variance (within + between) of the outcome variable.

For addressing Research Question 2, we used a doubly latent approach with cross-level measurement invariance for the construct of feeling of belonging with one's peers. This approach comprises a twofold procedure using latent variables: (a) latent measurement models at both levels and (b) latent aggregation for

the classroom level construct (e.g., Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Marsh et al., 2012). An important advantage of a model with these features is that it corrects possible measurement and sampling errors associated with designs in which variables measured at the individual level are used to operationalize a construct at the classroom level. We focused in our analyses on feeling of belonging with one's peers at the classroom level. However, we conducted an additional cross-level interaction analysis in a random slope multilevel structural equation model for Research Question 2a on the association between individual ethnic minority versus majority status and the relationship between the proportion of minority students and feeling of belonging with one's peers.

**Missing data.** Students in the sampled schools were obliged to participate in the achievement tests. Yet, individual students could be excluded from the study by the school if they met one of the following three criteria: (a) students with permanent physical impairment that made it impossible to participate, (b) severe intellectual or emotional impairment, and (c) students who were less than 1 year in Germany and could neither speak nor read in German. The response rate of the student questionnaire in the total sample was 87.3%; that is, it was lower than the response rate of achievement tests of 98.3% because participation was not mandatory in some federal states of Germany. The rate of the questionnaire varied between 76% in Hamburg and 98% in Hesse. The response rate for the parents' questionnaires was 81.4%.

We excluded special needs schools ( $N = 51$  schools), schools from the former GDR ( $N = 398$  schools), and classrooms with a large number of missing values concerning ethnic background information ( $>50\%$ ;  $N = 17$  schools) from the analyses. This led to a sample size of 19,457 students in 908 schools. The remaining sample did not include any missing values on the ethnic background variable because we classified missing information as unidentifiable and kept the student in the data set. Information on student gender was missing for 0.64% of the sample, on cognitive abilities for 6.91%, on mathematics achievement for 4.75%, on reading achievement for 4.77%, on SES for 29.62%, and on all four belonging items for 19.27% of the sample. To deal with item nonresponse, we used the full information maximum likelihood (FIML) estimator implemented in Mplus for all variables except student gender. This estimator applies a model-based approach to missing data (see Enders, 2010), using all information available from the model variables to estimate the model parameters. In doing so, we were able to use 96.43% of the intended sample. A total of 3.57% of the students had either missing gender information or missing values on all estimated variables (SES, cognitive abilities, and outcome variables) and was excluded during the analyses.

## Results

### Preliminary Analyses

**Comparison and selection of diversity measures.** First, we computed the composition and diversity measures (see “Classroom level independent and background variables” and Table S.1). Inspection of bivariate Pearson correlations at the classroom level (see Table S.2 available as online supplemental material) showed that the measures of ethnic diversity are highly correlated with each other and that they are also highly correlated with the com-



monly used “majority-minority approach”; that is, the proportion of minority students in a classroom (first column in Table S.2).

To gain further insight into the relationship between the proportion of minority students in a classroom and the diversity measures, we plotted the proportions of ethnic minority students for each classroom and the values for Simpson’s  $D$  as an example (Figure 1). The two measures are directly dependent on one another: In classrooms with very low proportions of minority students and consequently very high proportions of majority students, ethnic diversity is lower. In classrooms with less or equal to 50% of minority students, the correlation between the proportion of minority students and Simpson’s  $D$  is  $r = .99$  ( $p < .01$ ), and in classrooms with a proportion of minority students greater than 50% it is  $r = .33$  ( $p < .01$ ). That is, classrooms with high proportions of minority students vary in their ethnic heterogeneity. The correlation pattern for other diversity measures was comparable.

Because the high intercorrelations between the proportion of minority students in a classroom and the various measures of diversity may cause problems of multicollinearity and make it difficult to disentangle associations of variables with these two classroom characteristics, we decided to calculate the diversity measures without counting the proportion of majority students—that is German students—as a category. These measures thus depict the diversity of the proportion of ethnic minority students. Both measures together, the proportion of minority students and such a diversity index, describe the overall diversity of students in a classroom. This approach shows similar results to analyses that are based on a reduced sample of classrooms with a proportion of minority students greater than 50% but simultaneously allows us to use the complete sample with its greater power. The intercorrelations among the adapted measures as well as their bivariate correlation with the independent and outcome variables can be found in Table S.3 in the online supplemental material. These analyses indicate that the correlation patterns for the various measures of ethnic diversity are very similar to each other. For reasons of parsimony, we only show the main analyses for three exemplary

measures of diversity in the result section of this paper (for results of further analyses with the remaining measures, see Tables S.7 to S.9, available in the online supplemental material).

**Measurement model fit and variance of outcome variables between classrooms.** For the feeling of belonging with one’s classmates as an outcome, we first explored the fit of the doubly latent model with cross-level measurement invariance which showed acceptable model fit ( $\chi^2 = 379.680$ ,  $df = 7$ ,  $p < .05$ , root mean square error of approximation [RMSEA] = .058, Comparative Fit Index [CFI] = .951, standardized root mean square residual [SRMR]<sub>within</sub> = .037, SRMR<sub>between</sub> = .041). Such an unconditioned model without any predictors provides information on the amount of variance at both levels necessary to compute the intraclass correlation (ICC; variance at individual level = 0.31,  $SE = 0.01$ ; variance at classroom level = 0.03,  $SE = 0.003$ ; both variances significantly different from zero). The ICC estimates the proportion of the total variance that can be attributed to differences between classrooms. If there was no variation in the feeling of belonging with one’s classmates between classrooms, multilevel analyses would not be meaningful. In our study the proportion of variance between classrooms was 9%. As is commonly the case for noncognitive variables, this proportion is lower than the variation typically found for achievement between classrooms (see Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006). This was also true for the current analyses—for mathematics achievement, the proportion of variance between classrooms was 22% and for reading achievement it was 20%.

### Relationship Between the Proportion of Ethnic Minority Students, Ethnic Diversity, and Individual Student Outcomes

**Proportion of ethnic minority students in a classroom and student achievement.** The first set of our multilevel analyses explored the relationship between measures of the ethnic makeup of classrooms and students’ individual achievement in mathematics and reading comprehension. The result pattern of the classroom level variables is shown in Table 3 for mathematics and in Table 4 for reading comprehension as an outcome (for individual student-level results, see Table S.4 and S.5 in the online supplemental material). In a first step (Model M.1 in Table 3 and Model R.1 in Table 4, see Research Question 1a), we investigated the association between the proportion of ethnic minority students and achievement controlling for individual students’ ethnic background, gender, cognitive abilities, and SES. The results show that—independent of individual background characteristics—a student in a classroom with a one standard deviation higher proportion of ethnic minority students reached mathematics scores that were 0.17 SDs lower on average than a student in a class with a low proportion of ethnic minority students (Model M.1 in Table 3). Furthermore, the significant regression weight of the quadratic term of the proportion of ethnic minority students in Model M.1b indicates a slightly inverse U-shaped relationship with a tendency of lower mathematics scores in classrooms with very low or very high proportions of ethnic minority students. The same result pattern emerged for reading comprehension as an outcome (see Model R.1 and R.1b in Table 4).

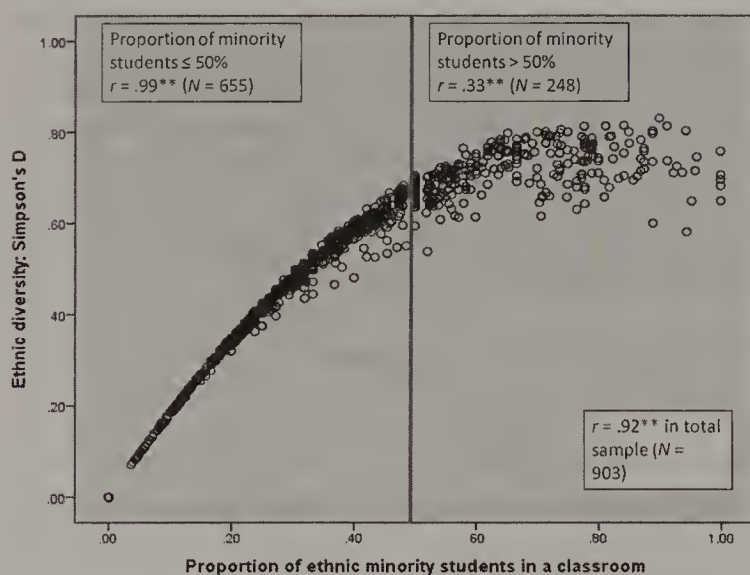


Figure 1. Joint distribution of the proportion of ethnic minority students in a classroom and Simpson’s  $D$  for  $N = 903$  classrooms. (Students with only German background, i.e., majority students, were counted as one ethnic group when calculating Simpson’s  $D$ ).



Table 3

Results of Multilevel Structural Equation Models Predicting Mathematics Achievement (Classroom-Level Results)

Variable	Model M.1	Model M.1b	Model M.2 (Ncat)	Model M.3 (SimD)	Model M.4 (ShaH)	Model M.5 (Ncat)	Model M.6 (SimD)	Model M.7 (ShaH)	Model M.8 (Ncat)	Model M.9 (SimD)	Model M.10 (ShaH)
Minority students %	-.17**	-.16**	—	—	—	-.19**	-.18**	-.18**	-.15**	-.14**	-.14**
Minority students %: Quadratic term	—	-.04**	—	—	—	-.03	-.03	-.03	-.00	-.00	-.00
Diversity measure	—	—	-.08**	-.05**	-.06**	.04*	.04*	.05*	.04*	.03*	.03*
Cognitive abilities (M)	—	—	—	—	—	—	—	—	.07**	.07**	.07**
SES (M)	—	—	—	—	—	—	—	—	.06**	.06**	.06**
R <sup>2</sup> L2	.25	.26	.05	.02	.03	.27	.27	.27	.39	.39	.39
R <sup>2</sup> L1	.35	.35	.35	.35	.35	.35	.35	.35	.34	.34	.34

Note. Ncat = number of categories (i. e. ethnicities); SimD = Simpson's D; ShaH = Shannon's H; SES = socioeconomic status; L1 = student level; L2 = classroom level. "Diversity measure" shows coefficients of the respective measure indicated in the first line. Covariates at the student level: ethnic background, SES, cognitive abilities, and gender. Regression coefficients were standardized by the total variance (within + between) of the outcome variable. For student-level results, see Table S.4, available as online supplemental material.

\*  $p < .05$ . \*\*  $p < .01$ .

Table 4

Results of Multilevel Structural Equation Models Predicting German Reading Achievement (Classroom-Level Results)

Variable	Model R.1	Model R.1b	Model R.2 (Ncat)	Model R.3 (SimD)	Model R.4 (ShaH)	Model R.5 (Ncat)	Model R.6 (SimD)	Model R.7 (ShaH)	Model R.8 (Ncat)	Model R.9 (SimD)	Model R.10 (ShaH)
Minority students %	-.15**	-.14**	—	—	—	-.13**	-.13**	-.13**	-.10**	-.10**	-.10**
Minority students %: quadratic term	—	-.04**	—	—	—	-.04*	-.04*	-.04*	-.01	-.01	-.01
Diversity measure	—	—	-.08**	-.06**	-.07**	-.01	-.00	-.01	.00	.00	.00
Cognitive abilities (M)	—	—	—	—	—	—	—	—	.05**	.05**	.05**
SES (M)	—	—	—	—	—	—	—	—	.07**	.07**	.07**
R <sup>2</sup> L2	.21	.23	.07	.03	.05	.23	.23	.23	.30	.30	.30
R <sup>2</sup> L1	.26	.26	.26	.26	.26	.26	.26	.26	.25	.25	.25

Note. Ncat = number of categories (i. e. ethnicities); SimD = Simpson's D; ShaH = Shannon's H; SES = socioeconomic status; L1 = student level; L2 = classroom level. "Diversity measure" shows coefficients of the respective measure indicated in the first line. Covariates at the student level: ethnic background, SES, cognitive abilities, and gender. Regression coefficients were standardized by the total variance (within + between) of the outcome variable. For student-level results, see Table S.5, available as online supplemental material.

\*  $p < .05$ . \*\*  $p < .01$ .

**Ethnic diversity in a classroom and student achievement.** The next models (M.2 to M.4 in Table 3 and R.2 to R.4 in Table 4) show results for selected diversity measures each as single predictor at the classroom level (see Research Question 1b). The first line of the result tables indicates the respective diversity measure used (Ncat = number of ethnicities in the classroom, SimD = Simpson's *D*, ShaH = Shannon's *H*; see Table S.7 to S.9 in the online supplemental material for analyses with further diversity measures). Their coefficients are slightly smaller in size than the coefficient for the proportion of minority students in the classroom, but they also show a negative relationship with individual mathematics and reading achievement.<sup>4</sup> That is, in ethnically more diverse classrooms, students show slightly lower levels of mathematics and reading achievement scores if associations with further classroom background characteristics are not controlled.

**Relationship between the proportion of ethnic minority students, ethnic diversity, and student achievement.** When the proportion of ethnic minority students in a classroom and a diversity measure are analyzed simultaneously as predictors at the classroom level (Models M.5 to M.7 in Table 3 and Models R.5 to R.7 in Table 4; see Research Question 1c) and also controlling for the socioeconomic composition and cognitive abilities level (Models M.8 to M.10 in Table 3 and Models R.8 to R.10 in Table 4), the coefficients for the proportion of ethnic minority students remain negative and significant for mathematics achievement and reading comprehension as outcome. However, the associations between ethnic diversity in the classroom and student achievement are different from the single predictor models: After taking into account the proportion of minority students and the classroom composition with regard to cognitive abilities and SES, respectively, ethnic diversity was slightly positively related to individual mathematics achievement (Models M.5 to M.10 in Table 3).<sup>5</sup> The identical analyses using reading achievement as an outcome (Models R.5 to R.10 in Table 4) showed no significant association with ethnic diversity. In all analyses, the proportion of explained variance did not differ much between models using various measures of ethnic diversity.

**Relationship between the proportion of ethnic minority students, ethnic diversity, and students feeling of belonging with their peers.** Analogous to the models presented so far, Table 5 shows the results for analyses exploring students' feeling of belonging with their peers as an outcome. The models using the proportion of ethnic minority students or a diversity measure as single predictor at the classroom level controlling for individual student characteristics showed negative relationships with students' feeling of belonging with their peers (Models B.1 to B.4; see Research Questions 2a and 2b). The strongest predictor was the proportion of ethnic minority students in a classroom. Its nonsignificant quadratic term showed that the relationship between the proportion of ethnic minority students and feeling of belonging was linear (Model B.1b).

Partly in line with assumptions mentioned regarding research question 2a, additional cross-level interaction analyses (Model B.1c in Table 5) revealed that individual ethnic majority students felt a stronger sense of belonging with their peers in classrooms with a higher proportion of ethnic majority students (interaction term:  $\beta = .23$ ,  $SE = .08$ ,  $p < .01$ ). For ethnic minority students the expected pattern did not emerge. Minority students on average

showed a weaker feeling of belonging with their peers than majority students did and their feeling of belonging with the classmates did not vary substantially depending on the proportion of minority students. In Models B.2 to B.4, the regression coefficients of the diversity measures were rather small and reached statistical significance only in some models. However, these measures represent diversity within the group of ethnic minority students. Models using diversity measures including German students as one category (not presented in Table 5) partly showed stronger associations (for Ncat:  $\beta = -.04$ ,  $SE = .01$ ,  $p < .01$ ; for SimD:  $\beta = -.07$ ,  $SE = .01$ ,  $p < .01$ ; for ShaH:  $\beta = -.07$ ,  $SE = .01$ ,  $p < .01$ ). When analyzed simultaneously as predictors at the classroom level (Models B.5 to B.7; see Research Question 2c) and also including classroom level covariates (Models B.8 to B.10), the coefficients of the proportion of ethnic minority students remained negative and significant and ethnic diversity within the group of minority students was not significantly related to students' feeling of belonging with their peers. In a last step, we explored the relationship between the ethnic makeup of classrooms, feeling of belonging with one's classmates, and student achievement in a full model (see Table S.10 for mathematics and Table S.11 for reading comprehension as an outcome). Students' feeling of belonging with their classmates was positively associated with achievement outcomes (for instance Model M.8b:  $\beta_{\text{mathematics}} = .11$ ,  $SE = .02$ ,  $p < .01$ , Model R.8b:  $\beta_{\text{reading}} = .07$ ,  $SE = .02$ ,  $p < .01$ ), and there was a slight indirect relationship between the proportion of minority students, feeling of belonging, and achievement outcomes (for instance, Model M.8b:  $\beta_{\text{mathematics}} = -.01$ ,  $SE = .00$ ,  $p < .01$ , Model R.8b:  $\beta_{\text{reading}} = -.01$ ,  $SE = .00$ ,  $p = .01$ ).

## Discussion

The present study investigated the relationship between various measures of ethnic composition and heterogeneity or diversity used in different disciplines on the one hand and achievement and psychosocial student outcomes on the other hand. The aim was to explore whether measures of ethnic diversity are related to student outcomes over and above commonly investigated characteristics of classroom composition. We sought to shed light on the question of whether the proportion of minority students is sufficient to investigate associations between different outcomes and the ethnic makeup of classrooms in terms of diversity or whether additional measures are necessary.

First, we therefore collected detailed information on possible diversity measures from research conducted in disciplines such as communication, geography, and biology. Our preliminary analyses comparing these measures operationalizing the ethnic makeup of

<sup>4</sup> These indices represent diversity among minority students (see section "Comparison and selection of diversity measures"). When the indices are calculated including German students as one group, the coefficients for mathematics achievement as an outcome are: Ncat:  $\beta = -.07$ ,  $SE = .01$ ,  $p < .01$ ; SimD:  $\beta = -.14$ ,  $SE = .01$ ,  $p < .01$ ; ShaH:  $\beta = -.13$ ,  $SE = .01$ ,  $p < .01$  and for reading achievement as an outcome they are: Ncat:  $\beta = -.08$ ,  $SE = .01$ ,  $p < .01$ ; SimD:  $\beta = -.13$ ,  $SE = .01$ ,  $p < .01$ ; ShaH:  $\beta = -.12$ ,  $SE = .01$ ,  $p < .01$ .

<sup>5</sup> The predictors at the classroom level were not highly interrelated, thus multicollinearity was not a concern (Variance inflation factor [VIF] for proportion of minority students = 1.53, for Simpson's *D* = 1.38, for average prior achievement = 1.32, and for average SES = 1.27). For correlation tables, see online supplemental Table S.3.



Table 5  
Results of Multilevel Structural Equation Models Predicting Feeling of Belonging With One's Peers (Classroom Level Results)

Variable	Model B.1	Model B.1b	Model B.1c	Model B.2 (Ncat)	Model B.3 (SimD)	Model B.4 (ShaH)	Model B.5 (Ncat)	Model B.6 (SimD)	Model B.7 (ShaH)	Model B.8 (Ncat)	Model B.9 (SimD)	Model B.10 (ShaH)
Minority students %	-.09**	-.08**	-.10**	—	—	—	-.10**	-.10**	-.10**	-.07**	-.07**	-.08**
Minority students %: quadratic term	—	-.01	—	—	—	—	-.00	.00	.00	.00	.01	.01
Diversity measure	—	—	—	-.04**	-.02	-.03*	.02	.02	.02	.01	.02	.02
Cognitive abilities (M)	—	—	—	—	—	—	—	—	—	.08**	.08**	.08**
SES (M)	—	—	—	—	—	—	—	—	—	-.02	-.02	-.02
Interaction: minority students % × individual minority background	—	—	.23**	—	—	—	—	—	—	—	—	—
R <sup>2</sup> L2	.11	.12	— <sup>a</sup>	.03	.01	.01	.12	.12	.12	.19	.19	.19
R <sup>2</sup> L1	.03	.03	— <sup>a</sup>	.03	.03	.03	.03	.03	.03	.03	.03	.03

Note. Ncat = number of categories (i. e. ethnicities); SimD = Simpson's D; ShaH = Shannon's H; SES = socioeconomic status; L1 = student level; L2 = classroom level. "Diversity measure" shows coefficients of the respective measure indicated in the first line. Covariates at the student level: ethnic background, SES, cognitive abilities, and gender. All regression coefficients (except for the interaction term) were standardized by the total variance (within + between) of the outcome variable. For student level results, see Table S.6, available as online supplemental material.

<sup>a</sup> Explained variance is not reported for random slope models, because in these models the variance of the outcome variable varies as a function of the predictor variables.

\*  $p < .05$ . \*\*  $p < .01$ .

classrooms led to the conclusion that they are highly intercorrelated (see McDonald & Dimmick, 2003), which is why we considered a selection of diversity measures within our main analyses. Overall, using different diversity measures as independent variable predicting student outcomes led to comparable results. Furthermore, the proportion of minority students—which is mostly used in educational research to describe the ethnic makeup—was highly correlated with diversity measures of the complete student population in a classroom. Thus, even if diversity and proportion of minority students are not the same from a content perspective, they should lead to comparable associations with student outcomes. In our study, we adapted the diversity measures to represent diversity only among minority students.

Our main analyses showed that students reached lower achievement scores in classrooms with a higher proportion of ethnic minority students (Hypothesis 1a). These findings are in line with international research (Mickelson et al., 2013; Van Ewijk & Sleegers, 2010a). Possible mediating and additional influential factors—which were not the focus of the present study—that may induce such findings are instructional quality, motivational processes among peers, non-German language spoken with peers, and school resources such as organizational structures and teacher competencies (Agirdag et al., 2012; Palardy, 2015; Raudenbush et al., 1998; Stipek, 2004; Van Ewijk & Sleegers, 2010a). Similarly, students reached lower achievement scores in ethnically more diverse classrooms. This result is in line with our hypothesis (see Hypothesis 1b, cf. Byrnes & Miller-Cotto, 2016) but contradicts studies showing an advantage for academic achievement development in ethnically diverse classrooms or schools (e.g., Tam & Bassett, 2004). However, when the proportion of ethnic minority students and a diversity measure were analyzed as joined predictors, controlling for average SES and average cognitive abilities in the classroom, we found different patterns of results: After accounting for differences in the proportion of minority students, ethnic diversity in the classroom was not significantly related to reading achievement, but did show a weak positive association with students' level of mathematics achievement. The findings are partly in line with assumptions of advantaged achievement development in ethnically diverse classrooms (Benner & Crosnoe, 2011; Gurin et al., 2003; Tam & Bassett, 2004). Thus, a positive association with mathematics achievement became visible only after controlling for the proportion of minority students, the level of socioeconomic status and cognitive abilities at the classroom level. We assume that controlling for these classroom characteristics also controlled for the less favorable learning environment and cumulated disadvantages in classrooms that are associated with a higher proportion of ethnic minority students.

The question arises why this pattern of slightly positive associations between diversity and achievement outcomes—which cannot be interpreted as causal relationships—emerged only for mathematics achievement. The reliability of the reading comprehension measure used was lower than that of the mathematics measure which might influence the findings. It is possible that the theoretically assumed benefits of ethnic classroom diversity develop easily for mathematics achievement because it is strongly tied to instruction (Crosnoe et al., 2010). As an alternative, it could be the case that a positive effect of ethnic diversity emerges for a large number of achievement outcomes but did not for reading comprehension because reading was tested in



German and ethnic minority students might speak not German at home. In addition, we can assume that in ethnically diverse classrooms there is a larger number of different language backgrounds present which might be related to different kinds of student difficulties and also different cultural background knowledge needed to understand texts. In such classrooms, it might be more difficult for the teacher to react to all students' needs during language instruction, exacerbating positive diversity effects. Future studies should explore additional explanations for positive associations between mathematics achievement and diversity. For instance, it is possible that diverse classrooms are also characterized by further beneficial features, such as school resources in terms of learning materials, teacher competencies and cultural understanding, organizational structure, and pedagogical concepts (cf. Eccles & Roeser, 2011). In the end, the current results do not allow the overall conclusion that ethnic diversity per se is beneficial for math learning. The weak positive associations found do not allow a causal interpretation in our cross-sectional design. They could emerge as a result of the former mechanisms but also because of unobserved characteristics of the classrooms in the present study or they could be related to the amount of missing ethnic background information and a classification as students with "unidentifiable" ethnic background. A replication of the result pattern in countries with another ethnic composition is needed.

Our multilevel analyses taking students' feeling of belonging with one's peers as an outcome revealed that students felt less related to their classmates in classrooms with a high proportion of ethnic minority students and in ethnically more diverse classrooms. Additional analyses showed that individual majority students felt a stronger sense of belonging in classrooms with higher proportions of majority students while minority students on average reported a weaker sense of belonging with their peers independent of the proportion of minority students. This finding was partly in line with the belongingness perspective predicting higher sense of belonging in more homogeneous groups (Baumeister & Leary, 1995; Benner & Crosnoe, 2011; Byrne, 1971; Tajfel & Turner, 1986). It is interesting that the diversity measures were not more strongly related to the feeling of belonging than the proportion of ethnic minority students was. Studies analyzing ethnic composition and belonging commonly argue that students build their belonging based on their specific ethnic background (see Benner & Crosnoe, 2011) rather than on the broader distinction between ethnic minority and majority. The present study initially indicates that overall minority versus majority group membership may be influential for students' feeling of belonging with their peers. Because the feeling of belonging with one's peers is an important psychosocial outcome of schooling that is positively related to motivational student characteristics (cf. Goodenow, 1993; Kumar & Maehr, 2010), classroom characteristics that are prone to foster a feeling of belonging should be further addressed in future research. A further interesting question that goes beyond the scope of the main research questions of this study is whether the feeling of belonging with one's peers partially mediates effects of the ethnic makeup of classrooms on achievement outcomes (cf. Christenson et al., 2012). First supplementary analyses of these relationships in full models indicate a positive association between students' feeling of belonging with their peers and achievement outcomes, as well as a slight indirect association between the proportion of minority students, feeling of belonging, and

achievement outcomes. However, in cross-sectional designs it is impossible to determine the direction of these relationships and possible reversed effects (e.g., achievement affects feeling of belonging).

In conclusion, the current findings indicate that using the proportion of minority students or a diversity index as measure of the ethnic makeup of classrooms mostly led to comparable conclusions. Including a diversity index in addition to the proportion of minority students showed additional weak associations with mathematics achievement and no significant relationship with reading achievement and feeling of belonging with one's peers.

### Limitations and Future Research

The present study has six important limitations. First, we analyzed data from a cross-sectional design, which renders it impossible to make statements about the origins and further development of the classroom effects including the positive associations found between ethnic diversity and mathematics achievement. One important background characteristic of classrooms that determines future student achievement is the average prior achievement in a classroom. We included it as a covariate using cognitive ability scores as a proxy. These test scores were collected at the same time point as the outcome variables and may therefore lead to a bias underestimating compositional effects (see Duncan, Magnuson, & Ludwig, 2004). At the same time, it is possible that our analyses overestimated classroom level associations as the proportion of minority students in a classroom is also confounded, for instance, with less favorable residential environments and segregated areas in large cities (for a methodological discussion of compositional effects, see Harker & Tymms, 2004; Hauser, 1970). Effects of the ethnic composition usually are smaller and often lack statistical significance when controlling for prior achievement level and SES level in German studies (Dumont, Neumann, Maaz, & Trautwein, 2013). Future research thus should favor longitudinal designs and include a large range of context characteristics.

Second, we were lacking background information on the country of origin for some students and created the category "unidentifiable" to include them into the analyses in order to get a full picture of the complete classroom. This may have led to distorted estimations of classroom level associations over- or underestimating associations between outcome variables and the ethnic makeup as they are treated as one category besides other ethnic categories and we lack information on how homogeneous this group is. However, excluding all students with missing background information seems not an option because it might underestimate the variety. Future research should gain this kind of background information for instance from school reports available for every student to avoid missing data. Furthermore, similar studies in other countries with a different ethnic composition than Germany are needed to disentangle effects of diversity and specific group characteristics within a system to a larger degree.

Third, we only analyzed relationships between the ethnic makeup of classrooms and students' feeling of belonging with one's classmates, as well as their achievement in mathematics and reading. Including a variety of achievement measures that



are more likely to be related to the benefits of exposure to diversity, such as creative thinking and problem solving (see Gurin et al., 2003), could be a useful addition in future studies.

Fourth, we do not know how the students in our study perceived ethnic diversity in their classroom and if their perception mattered for their attachment to classmates and group formation (for examples of perceived team diversity in organizational psychology see Shemla, Meyer, Greer, & Jehn, 2014). Future research should involve students' point of view to a larger degree.

Fifth, there may be further student background characteristics, such as SES and prior achievement, jointly constituting diversity in addition to the ethnic background. Recent developments in organizational psychology investigating the alignment of multiple diversity attributes and creating a hypothetical dividing line between homogeneous groups ("faultline"; see Thatcher & Patel, 2012) could be a model for future educational research as well.

Finally, our study pictures only one aspect of ethnic diversity—that is, the distribution of students in a classroom according to their families' country of birth expressed in a number that quantifies the degree of diversity. Future studies should explore additional operationalizations of and means to investigate diversity, for instance including information on the similarity between different ethnic groups, investigating latent characteristics of subgroups and interaction effects between individual and group characteristics, and explore means to identify mechanisms behind diversity effects. We are aware that ethnicity and ethnic identity go far beyond these numbers and would like to encourage more quantitative and qualitative research in this domain based on a diversity of methods. The aim of the present study was to explore measures of the ethnic makeup of classrooms in large-scale assessment frameworks. Thus, it provides a basis for future research that is concerned with recommendations on school and classroom composition and ways to address it.

## Conclusion

The findings indicate that the ethnic makeup of classrooms matters for individual achievement and psychosocial outcomes. The ethnic diversity measures collected for this educational research study ended up being closely intertwined. Thus, using one measure or the other should lead to comparable results. The proportion of ethnic minority students showed the strongest relation with student outcomes but ethnic diversity revealed slightly different result patterns for some outcomes. While the proportion of ethnic minority students in a classroom was negatively associated with individual student outcomes, ethnic diversity was positively related to mathematics achievement after controlling for relevant classroom background characteristics associated with less favorable learning environments. However, conclusions should be drawn cautiously. The slightly positive relationship between diversity and mathematics achievement needs replication in future research as it can be influenced for instance by unobserved characteristics of participating student groups. Future research in the field of education should not ignore diversity measures completely. Depending on the research question, subgroup and school subject diversity measures may give us more insight into how the ethnic makeup of the classroom is related to student outcomes.

## References

- Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162–172. <http://dx.doi.org/10.1016/j.stueduc.2005.05.008>
- Adams, R. J., Wu, M. L., & Carstensen, C. H. (2007). Application of multivariate Rasch models in international large-scale educational assessments. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 271–280). New York, NY: Springer. [http://dx.doi.org/10.1007/978-0-387-49839-3\\_17](http://dx.doi.org/10.1007/978-0-387-49839-3_17)
- Agirdag, O., Van Houtte, M., & Van Avermaet, P. (2012). Why does the ethnic and socio-economic composition of schools influence math achievement? The role of sense of futility and futility culture. *European Sociological Review, 28*, 366–378. <http://dx.doi.org/10.1093/esr/jcq070>
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*, 497–529. <http://dx.doi.org/10.1037/0033-2909.117.3.497>
- Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus [School structure and the creation of differential environments for learning and development]. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit* (pp. 95–188). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften. [http://dx.doi.org/10.1007/978-3-531-90082-7\\_4](http://dx.doi.org/10.1007/978-3-531-90082-7_4)
- Benner, A. D., & Crosnoe, R. (2011). The racial/ethnic composition of elementary schools and young children's academic and socioemotional functioning. *American Educational Research Journal, 48*, 621–646. <http://dx.doi.org/10.3102/0002831210384838>
- Benner, A. D., Graham, S., & Mistry, R. S. (2008). Discerning direct and mediated effects of ecological structures and processes on adolescents' educational outcomes. *Developmental Psychology, 44*, 840–854. <http://dx.doi.org/10.1037/0012-1649.44.3.840>
- Betts, J. R., Rueben, K. S., & Danenberg, A. (2000). *Equal resources, equal outcomes? The distribution of school resources and student achievement in California*. Public Policy Institute of California. Retrieved from [http://www.ppic.org/content/pubs/report/r\\_200jbr.pdf](http://www.ppic.org/content/pubs/report/r_200jbr.pdf)
- Biemann, T., & Kearney, E. (2010). Size does matter: How varying group sizes in a sample affect the most common measures of group diversity. *Organizational Research Methods, 13*, 582–599. <http://dx.doi.org/10.1177/1094428109338875>
- Biswas, A., & Mandal, S. (2010). Descriptive measures for nominal categorical variables. *Statistics & Probability Letters, 80*, 982–989. <http://dx.doi.org/10.1016/j.spl.2010.02.012>
- Blau, P. M. (1977). *Inequality and heterogeneity: A primitive theory of social structure*. New York, NY: Free Press.
- Bovaird, J. A. (2007). Multilevel structural equation models for contextual factors. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 149–182). Mahwah, NJ: Erlbaum.
- Budescu, D. V., & Budescu, M. (2012). How to measure diversity when you must. *Psychological Methods, 17*, 215–227. <http://dx.doi.org/10.1037/a0027129>
- Byrne, D. (1971). *The attraction paradigm*. New York, NY: Academic Press.
- Byrnes, J. P., & Miller-Cotto, D. (2016). The growth of mathematics and reading skills in segregated and diverse schools: An opportunity-propensity analysis of a national database. *Contemporary Educational Psychology, 46*, 34–51. <http://dx.doi.org/10.1016/j.cedpsych.2016.04.002>
- Christenson, S. L., Reschly, A. L., & Wylie, C. (2012). *Handbook of research on student engagement*. New York, NY: Springer. <http://dx.doi.org/10.1007/978-1-4614-2018-7>



- Crosnoe, R., Morrison, F., Burchinal, M., Pianta, R., Keating, D., Friedman, S. L., . . . the Eunice Kennedy Shriver National Institute of Child Health and Human Development Early Child Care Research Network. (2010). Instruction, teacher-student relations, and math achievement trajectories in elementary school. *Journal of Educational Psychology*, 102, 407-417. <http://dx.doi.org/10.1037/a0017762>
- Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11, 227-268. [http://dx.doi.org/10.1207/S15327965PLI1104\\_01](http://dx.doi.org/10.1207/S15327965PLI1104_01)
- Den Brok, P., & Levy, J. (2005). Teacher-student relationships in multicultural classes: Reviewing the past, preparing the future. *International Journal of Educational Research*, 43, 72-88. <http://dx.doi.org/10.1016/j.ijer.2006.03.007>
- Dimmick, J. R., & McDonald, D. G. (2001). Network radio as oligopoly, 1926-1956: Rivalrous imitation and program diversity. *Journal of Media Economics*, 14, 197-212. [http://dx.doi.org/10.1207/S15327736ME1404\\_1](http://dx.doi.org/10.1207/S15327736ME1404_1)
- Dougherty, K. D., & Huyser, K. R. (2008). Racially diverse congregations: Organizational identity and the accommodations of differences. *Journal for the Scientific Study of Religion*, 47, 23-44. <http://dx.doi.org/10.1111/j.1468-5906.2008.00390.x>
- Driessen, G. (2002). School composition and achievement in primary education: A large-scale multilevel approach. *Studies in Educational Evaluation*, 28, 347-368.
- Dumont, H., Neumann, M., Maaz, K., & Trautwein, U. (2013). Die Zusammensetzung der Schülerschaft als Einflussfaktor für Schulleistungen [The effect of student body composition on academic achievement: International and national evidence]. *Psychologie in Erziehung und Unterricht*, 60, 163-183. <http://dx.doi.org/10.2378/peu2013.art14d>
- Duncan, G. J., Magnuson, K. A., & Ludwig, J. (2004). The endogeneity problem in developmental studies. *Research in Human Development*, 1, 59-80. <http://dx.doi.org/10.1080/15427609.2004.9683330>
- Eccles, J. S., & Roeser, R. W. (2011). Schools as developmental contexts during adolescence. *Journal of Research on Adolescence*, 21, 225-241. <http://dx.doi.org/10.1111/j.1532-7795.2010.00725.x>
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: New York: Guilford Press.
- Entwisle, D. R., & Alexander, K. L. (1994). Winter setback: The racial composition of schools and learning to read. *American Sociological Review*, 59, 446-460. <http://dx.doi.org/10.2307/2095943>
- Fager, E. W. (1972). Diversity: A sampling study. *American Naturalist*, 106, 293-310. <http://dx.doi.org/10.1086/282772>
- Federal Statistical Office Germany. (2012). *Bevölkerung und Erwerbstätigkeit. Bevölkerung mit Migrationshintergrund—Ergebnisse des Mikrozensus 2011* [Population and occupation. Population with an immigrant background—Results of the micro-census 2011]. Wiesbaden, Germany: Statistisches Bundesamt. Retrieved from <https://www.destatis.de>
- Ganzeboom, H. B. G. (2010, May). *A new international socio-economic index [ISEI] of occupational status for the International Standard Classification of Occupation 2008 [ISCO-08] constructed with data from the ISSP 2002-2007; with an analysis of quality of educational measurement in ISSP*. Paper presented at Annual Conference of International Social Survey Programme, Lisbon. Retrieved from <http://www.harryganzeboom.nl/Pdf/2010-Ganzeboom-ISEI08-ISSP-Lisbon-%28paper%29.pdf>
- Ganzeboom, H. B. G., De Graaf, P. M., Treiman, D. J., & De Leeuw, J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, 21, 1-56. [http://dx.doi.org/10.1016/0049-089X\(92\)90017-B](http://dx.doi.org/10.1016/0049-089X(92)90017-B)
- Goldsmith, P. A. (2004). Schools' racial mix, students' optimism, and the Black-White and Latino-White achievement gaps. *Sociology of Education*, 77, 121-147. <http://dx.doi.org/10.1177/003804070407700202>
- Goldsmith, P. R. (2011). Coleman revisited: School segregation, peers, and frog ponds. *American Educational Research Journal*, 48, 508-535. <http://dx.doi.org/10.3102/0002831210392019>
- Goodenow, C. (1993). Classroom belonging among early adolescent students: Relationships to motivation and achievement. *The Journal of Early Adolescence*, 13, 21-43. <http://dx.doi.org/10.1177/0272431693013001002>
- Gurin, P. Y., Dey, E. L., Gurin, G., & Hurtado, S. (2003). How does racial/ethnic diversity promote education? *The Western Journal of Black Studies*, 27, 20-29.
- Hall, M., & Tideman, N. (1967). Measures of concentration. *Journal of the American Statistical Association*, 62, 162-168. <http://dx.doi.org/10.1080/01621459.1967.10482897>
- Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement*, 15, 177-199. <http://dx.doi.org/10.1076/sesi.15.2.177.30432>
- Harrison, D. A., & Klein, K. J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *The Academy of Management Review*, 32, 1199-1228. <http://dx.doi.org/10.5465/AMR.2007.26586096>
- Hauser, R. M. (1970). Context and consex: A cautionary tale. *American Journal of Sociology*, 75, 645-664. <http://dx.doi.org/10.1086/224894>
- Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4.-12. Klassen, Revision (KFT 4-12+ R)* [Cognitive abilities test for grades 4-12, revision]. Göttingen, Germany: Hogrefe.
- Herfindahl, O. C. (1950). *Concentration in the U.S. steel industry*. Unpublished doctoral dissertation, Columbia University, New York.
- Hoppe, A., Fujishiro, K., & Heaney, C. (2014). Workplace racial/ethnic similarity, job satisfaction, and lumbar back health among warehouse workers: Asymmetric reactions across racial/ethnic groups. *Journal of Organizational Behavior*, 35, 172-193. <http://dx.doi.org/10.1002/job.1860>
- Johnson, M. K., Crosnoe, R., & Elder, G. H. (2001). Students' attachment and academic engagement: The role of race and ethnicity. *Sociology of Education*, 74, 318-340. <http://dx.doi.org/10.2307/2673138>
- Junge, K. (1994). Diversity of ideas about diversity measurement. *Scandinavian Journal of Psychology*, 35, 16-26. <http://dx.doi.org/10.1111/j.1467-9450.1994.tb00929.x>
- Kumar, R., & Maehr, M. L. (2010). Schooling, cultural diversity, and student motivation. In J. L. Meece & J. S. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 308-324). New York, NY: Routledge.
- Kvålseth, T. O. (1991). Note on biological diversity, evenness, and homogeneity measures. *Oikos*, 62, 123-127. <http://dx.doi.org/10.2307/3545460>
- Les, M., & Maher, C. (1998). Measuring diversity: Choice in local housing markets. *Geographical Analysis*, 30, 172-190. <http://dx.doi.org/10.1111/j.1538-4632.1998.tb00395.x>
- Liebersohn, S. (1969). Measuring population diversity. *American Sociological Review*, 34, 850-862. <http://dx.doi.org/10.2307/2095977>
- Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods*, 16, 444-467. <http://dx.doi.org/10.1037/a0024376>
- MacArthur, R. H. (1965). Patterns of species diversity. *Biological Reviews of the Cambridge Philosophical Society*, 40, 510-533. <http://dx.doi.org/10.1111/j.1469-185X.1965.tb00815.x>
- Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, 47, 106-124. <http://dx.doi.org/10.1080/00461520.2012.670488>
- McCann, K. S. (2000). The diversity-stability debate. *Nature*, 405, 228-233. <http://dx.doi.org/10.1038/35012234>



- McDonald, D. G., & Dimmick, J. (2003). The conceptualization and measurement of diversity. *Communication Research*, 30, 60–79. <http://dx.doi.org/10.1177/0093650202239026>
- Meyer, B. (in press). Team diversity: A review of the literature. In R. Rico (Ed.), *The Wiley Blackwell handbook of the psychology of teamwork and collaborative processes*. Chichester, United Kingdom: Wiley-Blackwell.
- Mickelson, R. A., Bottia, M. C., & Lambert, R. (2013). Effects of school racial composition on K-12 mathematics outcomes: A metaregression analysis. *Review of Educational Research*, 83, 121–158. <http://dx.doi.org/10.3102/0034654312475322>
- Muthén, L. K., & Muthén, B. O. (1998–2010). Mplus (Version 6.1) [Computer software]. Los Angeles, CA: Author.
- Niemiec, C. P., & Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom: Applying self-determination theory to educational practice. *Theory and Research in Education*, 7, 133–144. <http://dx.doi.org/10.1177/1477878509104318>
- Organisation for Economic Co-operation and Development. (2010). *PISA 2009 results: Overcoming social background: Equity in learning opportunities and outcomes* (Vol. II). <http://dx.doi.org/10.1787/9789264091504-en>
- Ogbu, J. U. (1987). Variability in minority school performance: A problem in search of an explanation. *Anthropology & Education Quarterly*, 18, 312–334. <http://dx.doi.org/10.1525/aeq.1987.18.4.04x0022v>
- Ogbu, J. U. (2004). Collective identity and the burden of “Acting White” in Black history, community, and education. *The Urban Review*, 36, 1–35. <http://dx.doi.org/10.1023/B:URRE.0000042734.83194.f6>
- Okullo, P., & Moe, S. R. (2012). Large herbivores maintain termite-caused differences in herbaceous species diversity patterns. *Ecology*, 93, 2095–2103. <http://dx.doi.org/10.1890/11-2011.1>
- Palardy, G. J. (2015). Classroom-based inequalities and achievement gaps in first grade: The role of classroom context and access to qualified and effective teachers. *Teachers College Record*, 117, 1–48.
- Peetsma, T., Van der Veen, I., Koopman, P., & Van Schooten, E. (2006). Class composition influences on pupils’ cognitive development. *School Effectiveness and School Improvement*, 17, 275–302. <http://dx.doi.org/10.1080/13803610500480114>
- Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures*. Oxford, England: Viking.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Advanced quantitative techniques in the social sciences, Vol. 1). Newbury Park, CA: Sage.
- Raudenbush, S. W., Fotiu, R. P., & Cheong, Y. F. (1998). Inequality of access to educational resources: A national report card for eighth-grade math. *Educational Evaluation and Policy Analysis*, 20, 253–267. <http://dx.doi.org/10.3102/01623737020004253>
- Rauer, W., & Schuck, K.-D. (2003). *Fragebogen zur Erfassung emotionaler und sozialer Schulerfahrungen von Grundschulkindern dritter und vierter Klassen (FEES 3–4)* [Questionnaire for the assessment of emotional and social school experiences of primary school students in grade three and four]. Göttingen, Germany: Hogrefe.
- Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers’ perceptions of young children’s cognitive abilities: The role of child background and classroom context. *American Educational Research Journal*, 48, 335–360. <http://dx.doi.org/10.3102/0002831210374874>
- Richter, D., Engelbert, M., Böhme, K., Haag, N., Hannighofer, J., Reimers, H., . . . Stanat, P. (2012). Anlage und Durchführung des Ländervergleichs [Realization of the sample-based state comparison]. In P. Stanat, H. A. Pant, K. Böhme, & D. Richter (Eds.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011* (pp. 85–102). Münster, Germany: Waxmann.
- Rjosk, C., Richter, D., Hochweber, J., Lüdtke, O., & Stanat, P. (2015). Classroom composition and language minority students’ motivation in language lessons. *Journal of Educational Psychology*, 107, 1171–1185. <http://dx.doi.org/10.1037/edu0000035>
- Sanders, H. L. (1968). Marine benthic diversity: A comparative study. *American Naturalist*, 102, 243–282. <http://dx.doi.org/10.1086/282541>
- Seaton, E. K., & Yip, T. (2009). School and neighborhood contexts, perceptions of racial discrimination, and psychological well-being among African American adolescents. *Journal of Youth and Adolescence*, 38, 153–163. <http://dx.doi.org/10.1007/s10964-008-9356-x>
- Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.
- Shemla, M., Meyer, B., Greer, L., & Jehn, K. A. (2014). A review of perceived diversity in teams: Does how members perceive their team’s composition affect team processes and outcomes? *Journal of Organizational Behavior*, 37, 89–106. <http://dx.doi.org/10.1002/job.1957>
- Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163, 688. <http://dx.doi.org/10.1038/163688a0>
- Stanat, P., Pant, H. A., Böhme, K., & Richter, D. (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011* [German and mathematics competencies of students at the end of fourth grade: Results of the IQB sample-based state comparison]. Münster, Germany: Waxmann.
- Stipek, D. (2004). Teaching practices in kindergarten and first grade: Different strokes for different folks. *Early Childhood Research Quarterly*, 19, 548–568. <http://dx.doi.org/10.1016/j.ecresq.2004.10.010>
- Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4, 707–719. <http://dx.doi.org/10.1098/rsif.2007.0213>
- Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Chicago, IL: Nelson-Hall.
- Tam, M. Y. S., & Bassett, G. W. (2004). Does diversity matter? Measuring the impact of high school diversity on freshman GPA. *Policy Studies Journal: The Journal of the Policy Studies Organization*, 32, 129–143. <http://dx.doi.org/10.1111/j.0190-292X.2004.00056.x>
- Teachman, J. D. (1980). Analysis of population diversity: Measures of qualitative variation. *Sociological Methods & Research*, 8, 341–362. <http://dx.doi.org/10.1177/004912418000800305>
- Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers’ expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology*, 99, 253–273. <http://dx.doi.org/10.1037/0022-0663.99.2.253>
- Thatcher, S. M. B., & Patel, P. C. (2012). Group faultlines: A review, integration, and guide to future research. *Journal of Management*, 38, 969–1009. <http://dx.doi.org/10.1177/0149206311426187>
- Thorndike, R. L., & Hagen, E. (1971). *Cognitive Abilities Test (CAT). Examiner’s manual*. Boston, MA: Multi-Level-Edition.
- Thorndike, R. L., & Hagen, E. (1993). *Form 5 CogAT. Norms booklet*. Chicago, IL: Riverside.
- Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, 98, 788–806. <http://dx.doi.org/10.1037/0022-0663.98.4.788>
- Van Ewijk, R., & Sleegers, P. (2010a). Peer ethnicity and achievement: A meta-analysis into the compositional effect. *School Effectiveness and School Improvement*, 21, 237–265. <http://dx.doi.org/10.1080/09243451003612671>
- Van Ewijk, R., & Sleegers, P. (2010b). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review*, 5, 134–150. <http://dx.doi.org/10.1016/j.edurev.2010.02.001>

Walsh, J. A., & Taylor, R. B. (2007). Predicting decade-long changes in community motor vehicle theft rates: Impact of structure and surround. *Journal of Research in Crime and Delinquency*, 44, 64–90. <http://dx.doi.org/10.1177/0022427806295552>

Winkelmann, H., van den Heuvel-Panhuizen, M., & Robitzsch, A. (2008). Gender differences in the mathematics achievements of German primary school students: Results from a German large-scale study. *ZDM Mathematics Education*, 40, 601–616. <http://dx.doi.org/10.1007/s11858-008-0124-x>

Zusho, A., Pintrich, P. R., & Cortina, K. S. (2005). Motives, goals, and adaptive patterns of performance in Asian American and Anglo American students. *Learning and Individual Differences*, 15, 141–158. <http://dx.doi.org/10.1016/j.lindif.2004.11.003>

Received December 2, 2015

Revision received December 9, 2016

Accepted December 12, 2016 ■

### Call for Nominations

The Publications and Communications (P&C) Board of the American Psychological Association has opened nominations for the editorships of the *Journal of Experimental Psychology: Animal Learning and Cognition*, *Neuropsychology*, and *Psychological Methods* for the years 2020 to 2025. Ralph R. Miller, PhD, Gregory G. Brown, PhD, and Lisa L. Harlow, PhD, respectively, are the incumbent editors.

Candidates should be members of APA and should be available to start receiving manuscripts in early 2019 to prepare for issues published in 2020. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. Self-nominations are also encouraged.

Search chairs have been appointed as follows:

- *Journal of Experimental Psychology: Animal Learning and Cognition*, Chair: Stevan E. Hobfoll, PhD
- *Neuropsychology*, Chair: Stephen M. Rao, PhD
- *Psychological Methods*, Chair: Mark B. Sobell, PhD

Candidates should be nominated by accessing APA's EditorQuest site on the Web. Using your browser, go to <https://editorquest.apa.org>. On the Home menu on the left, find "Guests/Supporters." Next, click on the link "Submit a Nomination," enter your nominee's information, and click "Submit."

Prepared statements of one page or less in support of a nominee can also be submitted by e-mail to Sarah Wiederkehr, P&C Board Editor Search Liaison, at [swiederkehr@apa.org](mailto:swiederkehr@apa.org).

Deadline for accepting nominations is Monday, January 8, 2018, after which phase one vetting will begin.



Acknowledgments

The Editor, Steve Graham, thanks the following Principal Reviewers who evaluated manuscripts for *Journal of Educational Psychology* between June 1, 2016 and September 7, 2017.

Lisa Bendixen	Jim Folkestad	Daniel Moos	James P. Selig
Sarah Bonner		Christian Mueller	Michael J. Serra
	Hunter Gehlbach		Sungok Serena Shim
Simona C. S. Caravita		Florrie Ng	Jessica J. Summers
Stephanie Cawthon	Tanner Jackson		
Scotty D. Craig		Jay Parkes	Gita Taasooobshirazi
	Jeffrey D. Karpicke		Jessica Toste
Bridget Dalton	Evelyn Kroesbergen	Doug Rohrer	
Sidney K. D’Mello		Rod D. Roscoe	Kimberly Vannest
		Amy Rouse	
Anastasia D. Elder	Gregory Arief Liem		

The Editor also thanks the following ad hoc reviewers who evaluated manuscript for *Journal of Educational Psychology* between June 1, 2016 and September 7, 2017.

Rakefet Ackerman*	Daniel Berry	Chei-Chang Chiou	Meixia Ding
Elizabeth Adams	Sylvia Beyer	Jason C. Chow	Joseph Santino D’Intino
Pooja K. Agarwal	Patrick Beymer*	Amy Claessens	Benoît Dompnier
Yusra Ahmed	Alpana Bhattacharya	Carrie Clark	Markus Dresel
Joyce Alexander	Karen Bierman	Courtney Clark*	Kyle Du*
Andreas Alexiou	Kevin R. Binning	Aidan Clerkin	Angela L. Duckworth
Jennifer L. Allen	Elizabeth Bjork	Julia Cohen	Carolyn Dufault
Laura Kristen Allen	Robert Bjork	David Coker	Alana Dulaney
James Ryan Alverson	Courtney King Blackwell	Rebekah Levine Coley	Hanna Dumont
Ryan Alverson	Angel Blanch	Lissa Conley	John Dunlosky
Ross Anderson*	Peter Boedeker*	Carol M. Connor	George DuPaul
Kenn Apel	Daniel Bolt	Nicole Conrad	Amanda M. Durik
Katrin Arens	Mimi Bong	Johnathan Cook	
Mikko Aro	Julie L. Booth	Danya Corkin	
Mark Ashcraft	Renske Bouwer	Dewey Cornell	Leslie Echols
Steven R. Asher	Natasha K. K. Bowen	Jennifer H. Corpus	Malte Elson*
Yigal Attali	Edward Brent	Daniel Corral	Tino Endres*
Abbey Auxter	Britta Bresina*	Catherine Costigan	Sothy Eng
Anton Axelsson	M. Anne Britt	Rebecca Covarrubias	Michelle Englund
Paul Ayres	Cyril Brom	Marcus Crede	
	Elizabeth R. Brown	Amy Crosson	
Martine Baars	Rupert Brown	James Cummins	Guillermo Farfan*
Andrew Bacher-Hicks	Roger Bruning.	Malcolm Cunningham	George Farkas
Heather J. Bachman	Kristen Bub		Thomas Farmer
Han Suk Bae	Dung Bui	Samantha G. Daley	Dale Farran
Drew H. Bailey	Richard Burns	Lavinia E. Damian	Ingrid G. Farreras
Robert L. Bangert-Drowns	Ruth Butler	Noor Al Dahhan*	David Feldon
Christina Barbieri	James Byrnes	Celine Darnon	Sarah Lynn Ferguson
Eli Barnett*		Robert Davies	Dieter Ferring
Alison E. Baroody	Aimee A. Callender	Sarah Davies*	Logan Fiorella
Dan Battey	Lindsey Cameron	William E. Davis	Frank Fischer
Sheri Bauman	Brett D. Campbell	Bieke de Fraine	Abraham Flanigan*
Michael D. Beck	Elizabeth A. Canning	Peter F. de Jong	Lot Fonteyne
Michael Becker	Manuel Cargnino	Fien De Smedt	Donald J. Foss
Jonathan Beckett	Shana K. Carpenter	Helene Deacon	Matthew E. Foster
Elizabeth R. Bell	Beth M. Casey	Régine Debrosse*	Anne Frenzel
Hefer Bembenutty	Juan Cristobal Castro-Alonso	Marielle C. Dekker	Karin S. Frey
Moti Benita	Hugh W. Catts	Carolyn A. Denton	Jeffrey Froh
Marvin Berkowitz	Soo Eun Chae	Mesmin Destin	Douglas Fuchs
Donte Bernard	Jessica Chan*	Colin DeYoung	Carrie J. Furrer
Valerie-D. Berner	Nick Chang*	Irene-Anna N. Diakidoy	Scott Robert. Furtwengler
Virginia Berninger	Ouhao CHEN	Theresa Dicke	Emily R. Fyfe

- |                        |                              |                        |                           |
|------------------------|------------------------------|------------------------|---------------------------|
| Joseph Calvin Gagnon   | Paige Johnson*               | Diego Armando Luna     | Rollanda E. O'Connor      |
| Brian Galla*           | Suzanne H. Jones             | Bazaldua               | Paul A. O'Keefe           |
| Richard Gallagher      | Steve Joordens               |                        | Eva Oesterlen*            |
| Pamela Garner          | Regina Jucks                 | Xin Ma                 | Arturo Jr. Olivarez       |
| Hanna Gaspard          |                              | David MacPhee          | Thierry Olive             |
| David Geary            | Cigdem Kagitecibasi          | Frank Manis            | Carol Booth Olson         |
| Richard Göllner        | Slava Kalyuga                | Gwen C. Marchand       | Daniel Oppenheimer        |
| Julia Gorges           | Sean H. K. Kang              | Peter Marschik         |                           |
| Adele E. Gottfried     | Sun-Mee Kang                 | Herb W. Marsh          | Steven C. Pan             |
| Kelly Grace*           | Steven Karau                 | Shantal Marshall       | Roberto H. Parada         |
| Samuel Greiff          | Hugh Kearns                  | Sandra Martin-Chang    | Babette Park              |
| Kevin Grimm            | Timothy Kelley*              | James Anthony Martinez | Bernadette Park           |
| Lisa Grimm             | Wing-Wah Ki                  | Delphine Martinot      | Min-Kyung Park*           |
| Carola Grunschel       | Michael J. Kieffer           | Andrew J. Mashburn     | Philip D. Parker          |
| Elizabeth Gunderson    | Stephen Sherif Killingsworth | Jamaal Matthews        | Rachel Part*              |
|                        | Ha Yeon Kim                  | Percival Matthews      | Elise T. Pas              |
| Douglas Hacker         | Sun-A Kim                    | Sarah McCarthy         | David Paunesku            |
| Andreas Hadjar         | Yanghee Kim                  | Barbara L. McCombs     | Reinhard Pekrun*          |
| Debbie L. Hahs-Vaughn  | Thomas A. Kindermann         | Meghan McCormick       | Peng Peng                 |
| Vernon C. Hall         | Alli Klapp                   | Dana McCoy             | Franziska Perels          |
| David Hallowell*       | Kathrin Klingsieck           | Sean McCrea            | Anthony C. Perez          |
| Jill V. Hamm           | Nidhi Kohli                  | Kathleen McDermott     | Charles Perfetti          |
| Bridget Hamre          | Svjetlana Kolic-Vehovec      | Dean P. McDonnell      | Katherine Perkins         |
| Hyemin Han             | Magdalena Kriebel*           | Matthew McLarnon       | Maximilian Pfof           |
| Laura Hanish           | Helga Krinzinger             | Deja McLean*           | Huy P. Phan               |
| Paul Hanselman         | Amy Kunkel                   | Leigh McLean           | Hector Ponce              |
| Nicole Hansen          | Oi-man Kwok                  | Jake McMullen          | Paul Poteat               |
| Judith M. Harackiewicz |                              | Danielle McNamara      | Sarah R. Powell           |
| Jeff Harring           | Arena C. Lam                 | Beth Meisinger         | Tiziana Pozzoli           |
| Brenna Hassinger-Das   | Silvia Siu-Yin Lam           | Xiangzhi Meng          | Anna Praetorius           |
| Martin Hassler         | H. Chad Lane                 | Douglas Mennin         | Rilana Prenger            |
| Jarkko Hautamaki       | Jonas W. B. Lang             | Diana J. Meter         | Caroline Pulfrey          |
| Cameron Hecht*         | Fani Lauermann               | Bonnie J. F. Meyer     | Cynthia Puranik           |
| Angela Heine           | Jason Lawrence               | Jeremy Miciak          | Aryn A. Pyke              |
| Klaus Helkama          | Joshua F. Lawrence           | Amori Yee Mikami       |                           |
| Lisa-Marie Henderson   | Sara C. Lawrence             | David I. Miller        | Geetha Ramani             |
| Maciel M. Hernández    | Rebecca Lazarides            | Gloria E. Miller       | Gerardo Ramirez           |
| Annemarie Hindman      | Campbell Leaper              | Keith Millis           | Kevin L. Rand             |
| Aline Hitti            | In Heok Lee                  | Kathi Miner            | Judy Randi                |
| Connie Suk Han Ho      | Jihyun Lee                   | Jens Moeller           | John Ranellucci           |
| Sarah Isabelle Hofer   | Kerry Lee                    | Ida Mok                | Luke Rapa                 |
| Adam Hoffman           | You-kyung Lee                | Anna Mueller           | Catherine F. Ratelle      |
| Solveig Holen          | James D. Lehman              | Kou Murayama           | Erik Rawls*               |
| Michelle Hood          | Christine Leider             | Mary Murphy            | Jenni Redifer             |
| Josefine Horbach       | Wolfgang Lenhard             | Maida Mustafic*        | Deborah K. Reed           |
| Robert Horner          | Che Kan Leong                | Aaron Myers            | Johnmarshall Reeve        |
| Lisette Hornstra       | Arne Lervag                  |                        | Robert D. Renaud          |
| Hayley Houseman*       | Tony Leung                   | William Nagy           | Alexander Renkl           |
| Londi Howard*          | Detlev Leutner               | Jennifer W. Neal       | Ilyse Resnick             |
| Jan N. Hughes          | Neil Lewis Jr                | Ron Nelson             | Jan Retelsdorf            |
| Darrell M. Hull        | Stephanie Lichtenfeld        | Tuan Nguyen            | David Rettinger           |
| Carol S. Huntsinger    | Karmela Liebkind             | Christopher Niemiec    | Matthew R. Reynolds       |
| Noelle Hurd            | Teresa Limpo                 | Markku Niemivirta      | Katherine T. Rhodes       |
| Michelle Hurst         | Li Liu                       | Christoph Niepel       | Luisa Ribeiro*            |
| Janet S. Hyde          | Gerine Lodder                | Yu Niiya               | Jessie Ricketts           |
| Jukka Hyönä            | Jason Micheal Lodge          | Adrienne Nishina       | Garrett Roberts           |
|                        | Abbey Loehr*                 | George Harvey Noell    | Robert Roeser             |
| Tiffany Ito            | Jessica Logan                | Elizabeth Norton       | Toni Rogat                |
|                        | Caitlin Lombardi             | Elena Novak            | Kathy Roskos              |
| Molly M. Jameson       | Christopher J. Lonigan       | Matthias Nückles       | Guy Roth                  |
| Jeremy Jamieson        | Mantou Lou                   | Jari-Eric Nurmi        | Deborah Wells Rowe        |
| Stefan Janke*          | Rebecca Lucas                |                        | Christine M. Rubie-Davies |
| Malte Jansen           | Steven Luke                  | Andreas Obersteiner    | Kathleen Moritz Rudasill  |
| Austin Jennings*       |                              | Erin O'Connor          | Nikol Rummel              |



Danielle Ruscio*	Jaehyun Shin*	Allen Thurston	Min Wang
Teomara Rutherford	Jiyun Elizabeth L. Shin	Simon Tiffin-Richards	Yanlin Wang*
Allison M. Ryan	Ken Shores	Liudmilla Titova*	Jeanne Wanzek
	Hua Shu	Tammy D. Tolar	Hersh Waxman
Chrissy Sandman*	Georgios D. Sideridis	Liliana Tolchinsky	Charles A. Weaver
Laila Sanguras	Johannes Siegrist	Ulrich Trautwein	Mi-young Lee Webb
Carol Sansone	Rebecca Silverman	Rebecca Treiman	Christina Weiland
Nicole Scalise*	Jessi L. Smith	Danijela Trenkic	Yana Weinstein
Blanca Schaefer	Kate Snyder	Adrea Truckenmiller	Allan Wigfield
Christopher Schatschneider	Nicolas Sommet		Daniel Willingham
Katerina Schenke	David Sparks	Patricia F. Vadasy	Melody Wiseheart
Sebastian Schmid	Jonathan Michael Spector	Huub van den Bergh	Phillip Karl Wood
Jennifer A. Schmidt	Rayne A. Sperling	Hans van der Meij	Trent Wondra*
Florian Schmiedek	Matthias Stadler	Martin Van Boekel*	Stephanie Virgine Wormington
Wolfgang Schnotz	Petra Stanat	Tamara van Gog	Wei Wu
Rob Schoonen	Jon R. Star	Peggy N. Van Meter	
Johannes Schult	Laura Steacy	Stefanie van Ophuysen	Jianzhong Xu
Ann Christine Schulte	Lori Stephens*	Mark Van Ryzin	
Matthias Schwaighofer*	Gabriele Steuer*	Nico Van Yperen	Ji Seung Yang
Dan Schwartz	Deborah Stipek	Jeffrey Vancouver	Ling-Yan Yang
Sarah Schwartz	Katherine Strasser	Maarten Vansteenkiste	Eunice Pui Yu Yim
Amy Schweinle	Andrew T. Stull	Marina Vasilyeva	Guo Ying
Colleen M. Seifert	Robert H. Stupnisky	Ariana Vasquez	Maaly Younis*
Katerina Sergi*	Amanda Sullivan	Dana Vedder-Weiss	Paula Yust*
Eran Shadach	Congying Sun*	Elizabeth Votruba-Drzal	
Shaul Shalvi	Michael I. Swart	Lien Vu*	Imac Maria Zambrana
Cynthia R. Shanahan	John Sweller	Heidi A. Vuletich	Dongbo Zhang
Lina Shanley			Jie Zhang
Rebecca J. Shearer	Zsofia Katalin Takacs	Tracy Evian Waasdorp	Qin Zhao
Ken Sheldon	Cheng Yong Tan	Chris Wahlheim	Jinxin Zhu
David Sherman	Roman Taraban	Jonathan Wai	Matthias Ziegler
Dara Shifrer	Monja Thiebach*	Huanhuan Wang*	

\* Denotes a reviewer who co-reviewed under the supervision of a primary reviewer.

**Call for Papers**  
**A Focused Collection of Qualitative Studies in the Psychological Sciences:  
Reasoning and Participation in Formal and Informal Learning Environments**

*Journal of Educational Psychology*

Guest Editors: Tanner LeBaron Wallace and Eric Kuo

Reasoning and participation are two central topics of education research in the psychological sciences. Understanding the mechanisms that govern thought and reasoning has long been a core enterprise of educational psychology and, over time, more modern views on learning have promoted participation as a key feature for research—either as a facilitator of learning, a practice to be learned, or as an operationalization of learning itself.

We are pleased to announce a focused collection highlighting qualitative studies of reasoning and participation in formal and informal learning environments. By inviting studies incorporating qualitative methods, we aim to complement the experimental and longitudinal statistical research on these topics that is typically published in this journal. We encourage submission of papers focused on the following (or closely related) topics:

- Student reasoning and/or participation in novel learning environments or activities
- The relations between student reasoning, motivation, identity, and participation
- Student perceptions and meaning-making during participatory experiences
- Dynamic models of student reasoning that are grounded in data
- Explanatory accounts for how and why participation is successful (or not)
- Identifying new goals or targeted outcomes for reasoning or participation

We especially welcome qualitative studies that demonstrate the possibilities for unique discovery afforded by inductive analysis of rich data sources (e.g., real-time recordings of student reasoning, participation, discourse, and physical action, students' meaning-making anchored to particular interactions experienced). This collection will highlight the benefits of qualitative methods for extending and deepening theoretical and empirical understandings of reasoning and participation in both formal and informal learning environments.

The deadline for manuscript submissions is **March 1, 2018**. We invite authors to contact the Guest Editors of this collection, Tanner LeBaron Wallace (twallace@pitt.edu) and Eric Kuo (erickuo@pitt.edu), for discussion on how to maximize alignment between their submissions and this focused collection, though it is not required. Please follow both APA guidelines as well as specific submission criteria for the journal. When submitting manuscripts, please also indicate your intent to submit to this focused collection in the required cover letter.

All manuscripts must be submitted electronically at <http://www.editorialmanager.com/edu>. In the submission portal, please select the article type "Special Section: Reasoning & Participation – Qualitative." For more information on the *Journal of Educational Psychology*, please visit <http://www.apa.org/pubs/journals/edu/>.



Instructions to Authors  
*Journal of Educational Psychology*  
www.apa.org/pubs/journals/edu

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

**Manuscript preparation.** Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (6th ed.). Manuscripts may be copyedited for bias-free language (see pp. 70–77 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see [www.apa.org/pubs/journals/edu](http://www.apa.org/pubs/journals/edu). **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychological Bulletin*, 139, 133–151. <http://dx.doi.org/10.1037/a0028566>

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Gill, M. J., & Sypher, B. D. (2009). Workplace incivility and organizational trust. In P. Lutgen-Sandvik & B. D. Sypher (Eds.), *Destructive organizational communication: Processes, consequences, and constructive ways of organizing* (pp. 53–73). New York, NY: Taylor & Francis.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see p. 34 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied in TIFF or EPS format. APA's policy on publication of color figures is available at <http://www.apa.org/pubs/authors/instructions.aspx?item=6>.

**Publication policies.** APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at [www.apa.org/pubs/authors/posting.aspx](http://www.apa.org/pubs/authors/posting.aspx). In addition, it is a violation of APA Ethical Principles to publish “as original data, data that have been previously published” (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in

whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that “after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release” (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

**Masked review policy.** The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., “in our previous work, Johnson et al., 1998 reported that . . .” Instead, references to the authors' work should be in third person, e.g., “Johnson et al. (1998) reported that . . .” The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at [www.apa.org/ethics/](http://www.apa.org/ethics/) or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

**Permissions.** Authors of accepted papers must obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including test materials (or portions thereof), photographs, and other graphic images (including those used as stimuli in experiments). On advice of counsel, APA may decline to publish any image whose copyright status is unknown.

**Supplemental materials.** APA can place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see [www.apa.org/pubs/authors/supp-material.aspx](http://www.apa.org/pubs/authors/supp-material.aspx) for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

**Submission.** Authors should submit their manuscripts electronically via the Manuscript Submission Portal at [www.apa.org/pubs/journals/edu/index.aspx](http://www.apa.org/pubs/journals/edu/index.aspx) (follow the link for submission under Instructions to Authors). General correspondence may be addressed to the editorial office at [CJohnson@apa.org](mailto:CJohnson@apa.org).

# CRITICAL THINKING ABOUT RESEARCH

PSYCHOLOGY AND RELATED FIELDS  
SECOND EDITION



To view the full Table of Contents,  
visit [www.apa.org/h4318149](http://www.apa.org/h4318149)

## Critical Thinking About Research

Psychology and Related Fields  
SECOND EDITION

*Julian Meltzoff and Harris Cooper*

To become informed consumers of research, students need to thoughtfully evaluate the research they read rather than accept it without question. This second edition of a classic text gives students what they need to apply critical reasoning when reading behavioral science research. It updates the original text with recent developments in research methods, including a new chapter on meta-analyses.

Part I gives a thorough overview of the steps in a research project. It focuses on how to assess whether the conclusions drawn in a behavioral science report are warranted by the methods used in the research. Topics include research hypotheses, sampling, experimental design, data analysis, interpretation of results, and ethics.

Part II allows readers to practice critical thinking with a series of fictitious journal articles containing built-in flaws in method and interpretation. Clever and engaging, each article is accompanied by a commentary that points out the errors of procedure and logic that have been deliberately embedded in the article. This combination of instruction and practical application will promote active learning and critical thinking in students studying the behavioral sciences. 2018. 541 pages. Paperback.

List: \$49.95 | APA Member/Affiliate: \$39.95 | ISBN 978-1-4338-2710-5 | Item # 4318149

TO ORDER: 800-374-2721 ▪ [www.apa.org/h4318149](http://www.apa.org/h4318149)

In Washington, DC, call: 202-336-5510 ▪ TDD/TTY: 202-336-6123 ▪ Fax: 202-336-5502

In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972



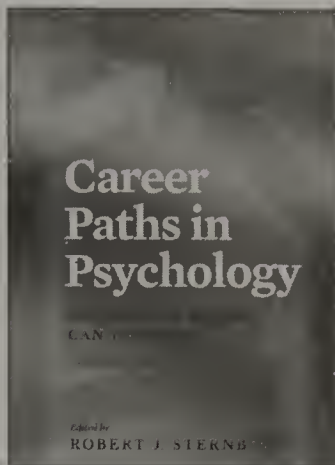


AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

# CAREER PATHS IN PSYCHOLOGY

Where Your Degree Can Take You  
THIRD EDITION

Edited by Robert J. Sternberg



Now in its third edition, this bestselling volume has set the standard for students seeking to find an exciting career in psychology. Its comprehensive coverage spans more careers than ever, with the vast majority of chapters new to this edition.

An advanced degree in psychology offers an extremely wide range of rewarding and well-compensated career opportunities. Amidst all the choices, this book will help future psychologists find their optimal career path. The chapters describe 30 different graduate-level careers (i.e., careers for those holding a PhD, EdD, or

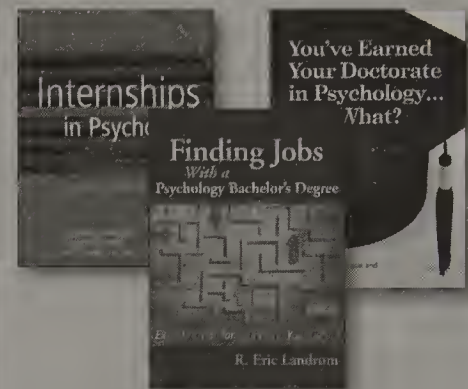
PsyD) in three distinct areas of endeavor: academia, clinical and counseling psychology, and specialized settings such as for-profit businesses, nonprofits, the military, and schools. Each chapter explores a different career, and describes typical daily activities, the approximate range of compensation, advantages and disadvantages of the career, opportunities for employment and advancement, and how to plan one's educational experiences to prepare for this specialty. The authors—all highly accomplished professionals—were selected for their years of experience, their distinction in their field, and their ability to communicate their passion. 2017. 584 pages. Paperback.

List: \$29.95 | APA Member/Affiliate: \$24.95 | ISBN 978-1-4338-2310-7 | Item # 4313041

## CONTENTS

Introduction | **I. ACADEMIA** | 1. Psychologists in University Departments of Psychology or Psychological Science | 2. Psychologists in College Departments of Psychology or Psychological Science | 3. Psychologists in Schools of Education | 4. Psychologists in Schools of Business | 5. Psychologists in Medical Schools | 6. Psychologists in Law Schools | 7. Psychologists in Schools of Public Policy | **II. CLINICAL AND COUNSELING PSYCHOLOGY** | 8. Clinical Psychologists in Independent Practice | 9. Psychologists Specializing in Child and Adolescent Clinical Psychology | 10. Geropsychologists: Psychologists Specializing in Aging | 11. Clinical Neuropsychologists | 12. Counseling Psychologists | 13. Psychologists Specializing in Psychopharmacology | 14. Psychologists Specializing in Rehabilitation Psychology | **III. SPECIALIZED SETTINGS** | 15. Psychologists Working in Independently Funded Research Centers and Institutes | 16. Forensic Psychologists | 17. Sport Psychologists | 18. Media Psychologist | 19. Consulting and Organizational Psychologists | 20. Psychologists in Management | 21. Consumer Psychologists | 22. Psychologists in the Publishing World | 23. Psychologists Writing Textbooks | 24. Military Psychologists | 25. Police and Public Safety Psychologists | 26. Psychologists Giving Grants Through Nonprofits | 27. Psychologists Giving Grants Through Government Organizations | 28. Psychologists in Educational Testing and Measurement Organizations | 29. School Psychologists | 30. Psychologists Pursuing Scientific Research in Government Service | Epilogue: Preparing for a Career in Psychology | Index | About the Editor

## ALSO OF INTEREST



### Internships in Psychology The APAGS Workbook for Writing Successful Applications and Finding the Right Fit

Carol Williams-Nickelson, Mitchell J. Prinstein, and W. Gregory Keilin  
2013. 120 pages. Paperback.

List: \$27.95 | APA Member/Affiliate: \$22.95  
ISBN 978-1-4338-1210-1 | Item # 4313034  
AVAILABLE ON AMAZON KINDLE®

### Finding Jobs With a Psychology Bachelor's Degree Expert Advice for Launching Your Career

R. Eric Landrum  
2009. 158 pages. Paperback.

List: \$24.95 | APA Member/Affiliate: \$19.95  
ISBN 978-1-4338-0437-3 | Item # 4313023  
AVAILABLE ON AMAZON KINDLE®

### You've Earned Your Doctorate in Psychology... Now What?

Securing a Job as an Academic  
or Professional Psychologist  
Elizabeth M. Morgan  
and R. Eric Landrum  
2012. 190 pages. Paperback.

List: \$24.95 | APA Member/Affiliate: \$24.95  
ISBN 978-1-4338-1145-6 | Item # 4313033  
AVAILABLE ON AMAZON KINDLE®

**APA BOOKS ORDERING INFORMATION: 800-374-2721 • www.apa.org/pubs/books**

In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502

In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD3114

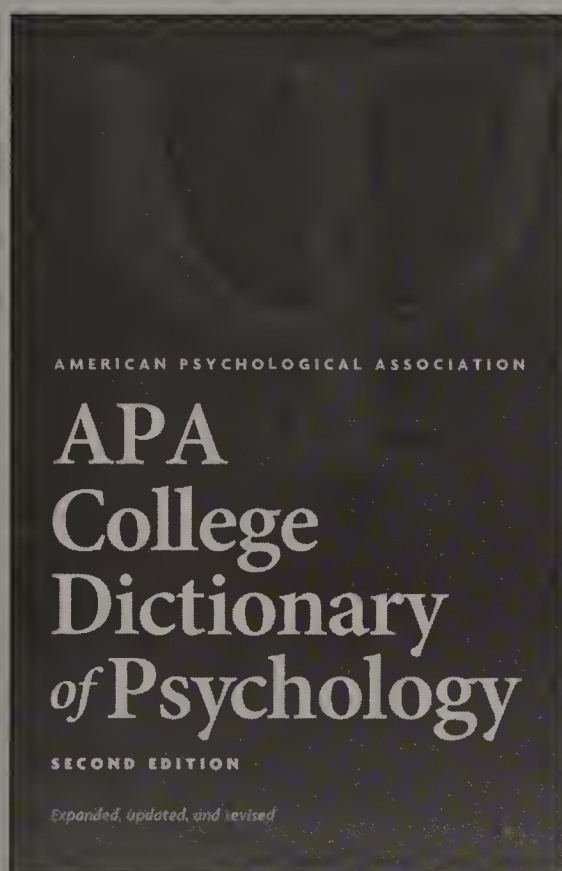


AMERICAN PSYCHOLOGICAL ASSOCIATION

# APA COLLEGE DICTIONARY OF PSYCHOLOGY

SECOND EDITION

Editor-in-Chief Gary R. VandenBos



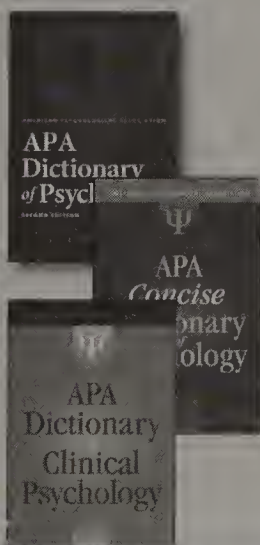
The American Psychological Association is excited to offer the second edition of its popular, compact, and economical student's dictionary. With some 5,500 entries—over 500 more than the original edition—the second edition continues to feature clear and authoritative definitions that provide basic coverage from across 90 subdisciplines of psychology.

Special emphasis is concentrated on the fields that are typically encountered in undergraduate studies, such as general, personality and social, lifespan developmental, abnormal, and cognitive psychology. Moreover, basic coverage of neuropsychology and of statistics and methodology have been enhanced for the second edition, and two helpful appendixes have been included: Abbreviations and Acronyms and Symbols.

*The APA College Dictionary of Psychology, Second Edition* is a reliable resource that answers the needs of both advanced placement high-school students and college undergraduates—whether they are taking psychology as part of a broader curriculum or making it their major field of study. 2016. 518 pages. Paperback.

List: \$19.95 | APA Member/Affiliate: \$19.95 | ISBN 978-1-4338-2158-5 | Item # 4311027

## ALSO OF INTEREST



A CHOICE OUTSTANDING  
ACADEMIC TITLE  
**APA Dictionary  
of Psychology**  
SECOND EDITION  
2015. 1,204 pages. Hardcover.

List: \$49.95  
APA Member/Affiliate: \$39.95  
ISBN 978-1-4338-1944-5  
Item # 4311022  
AVAILABLE ON AMAZON KINDLE®

**APA Concise  
Dictionary  
of Psychology**  
Editor-in-Chief  
Gary R. VandenBos  
2009. 583 pages. Hardcover.

List: \$39.95  
APA Member/Affiliate: \$29.95  
ISBN 978-1-4338-0391-8  
Item # 4311009  
AVAILABLE AS A MOBILE APP!

**APA Dictionary of  
Clinical Psychology**  
Editor-in-Chief  
Gary R. VandenBos  
2013. 636 pages. Hardcover.

List: \$39.95  
APA Member/Affiliate: \$29.95  
ISBN 978-1-4338-1207-1  
Item # 4311016  
AVAILABLE ON AMAZON KINDLE®

APA BOOKS ORDERING INFORMATION: 800-374-2721 • [www.apa.org/pubs/books](http://www.apa.org/pubs/books)

In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502

In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD3077



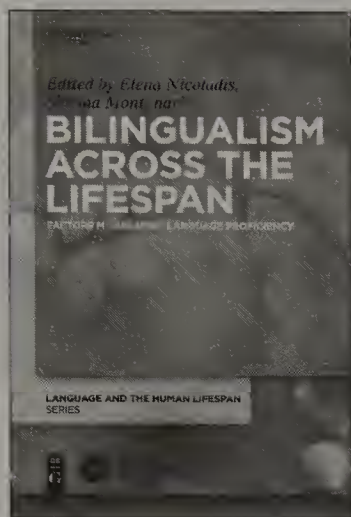


AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

# BILINGUALISM ACROSS THE LIFESPAN

Factors Moderating Language Proficiency

Edited by Elena Nicoladis and Simona Montanari



This book pioneers the study of bilingualism across the lifespan and in all its diverse forms. In framing the newest research within a lifespan perspective, the editors highlight the importance of considering an individual's age in researching how bilingualism affects language acquisition and cognitive development. A key theme is the variability among bilinguals, which may be due to a host of individual and sociocultural factors, including the degree to which bilingualism is valued within a particular context. Thus, this book is a call for language researchers,

psychologists, and educators to pursue a better understanding of bilingualism in our increasingly global society. 2016. 496 pages. Hardcover.

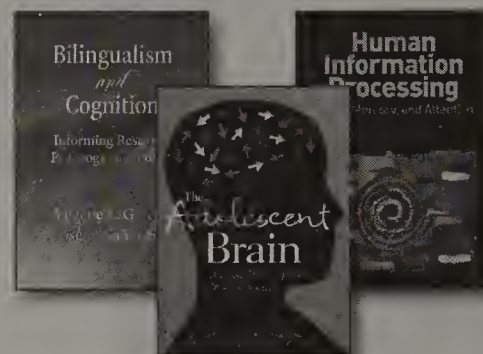
**Series: Language and the Human Lifespan**

List: \$79.95 | APA Member/Affiliate: \$49.95 | ISBN 978-1-4338-2283-4 | Item # 4316172

## CONTENTS

Introduction, *Simona Montanari and Elena Nicoladis* | Chapter 1. Shifting Perspectives on Bilingualism, *Fred Genesee* | **I. Early Bilingualism** | Chapter 2. Early Bilingualism: From Differentiation to the Impact of Family Language Practices, *Suzanne Quay and Simona Montanari* | Chapter 3. Speech Perception in Simultaneously Bilingual Infants, *Christopher T. Fennell, Angeline Sin-Mei Tsui, and Tamara M. Hudon* | Chapter 4. Early Lexical Development in Bilingual Infants and Toddlers, *Barbara T. Conboy and Simona Montanari* | Chapter 5. Code-Switching in Childhood, *W. Quin Yow, Ferninda Patrycia, and Suzanne Flynn* | **II. Factors Affecting Bilingualism Across the Lifespan** | Chapter 6. Quantity and Quality of Language Input in Bilingual Language Development, *Sharon Unsworth* | Chapter 7. Factors Moderating Proficiency in Bilingual Speakers, *Virginia C. Mueller Gathercole* | Chapter 8. Age of Onset of Bilingualism Effects and Availability of Input in First Language Attrition, *Silvina Montrul* | Chapter 9. Age of Second-Language Acquisition: Critical Periods and Social Concerns, *David Birdsong and Jan Vanhove* | Chapter 10. Code-Switching in Adulthood, *Jeff MacSwan* | **III. Academic Achievement and Literacy in Bilinguals** | Chapter 11. Bilingualism and Academic Achievement in Children in Dual Language Programs, *Kathryn Lindholm-Leary* | Chapter 12. Literacy in Adulthood: Reading in Two Languages, *Judith F. Kroll, Jason Gullifer, and Megan Zirnstein* | **IV. Cognitive Effects of Bilingualism** | Chapter 13. Cognitive Effects of Bilingualism in Infancy, *Ágnes Melinda Kovács* | Chapter 14. Bilingual Speakers' Cognitive Development in Childhood, *Elena Nicoladis* | Chapter 15. Cognitive and Emotional Effects of Bilingualism in Adulthood, *Max R. Freeman, Anthony Shook, and Viorica Marian* | Chapter 16. The Contribution of Bilingualism to Cognitive Reserve in Healthy Aging and Dementia, *Hilary D. Duncan and Natalie A. Phillips* | **V. Conclusion** | Concluding Remarks and Future Directions, *Simona Montanari and Elena Nicoladis* | Index | About the Editors

## ALSO OF INTEREST



### Bilingualism and Cognition Informing Research, Pedagogy, and Policy

Eugene E. García and José E. Náñez, Sr.  
2011. 242 pages. Hardcover.

List: \$39.95 | APA Member/Affiliate: \$34.95  
ISBN 978-1-4338-0879-1 | Item # 4318087

### The Adolescent Brain Learning, Reasoning, and Decision Making

Edited by  
Valerie F. Reyna, Sandra B. Chapman,  
Michael R. Dougherty, and Jere Confrey  
2012. 457 pages. Hardcover.

List: \$49.95 | APA Member/Affiliate: \$39.95  
ISBN 978-1-4338-1070-1 | Item # 4318098

AVAILABLE ON AMAZON KINDLE®

### Human Information Processing Vision, Memory, and Attention

Edited by  
Charles Chubb, Barbara A. Doshier,  
Zhong-Lin Lu, and Richard M. Shiffrin  
2013. 264 pages. Hardcover.

List: \$49.95 | APA Member/Affiliate: \$39.95  
ISBN 978-1-4338-1273-6 | Item # 4318115

AVAILABLE ON AMAZON KINDLE®



PsycBOOKS®

Access to chapters from a variety  
of APA scholarly & professional books.

APA BOOKS ORDERING INFORMATION: 800-374-2721 • [www.apa.org/pubs/books](http://www.apa.org/pubs/books)

In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502

In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD3093



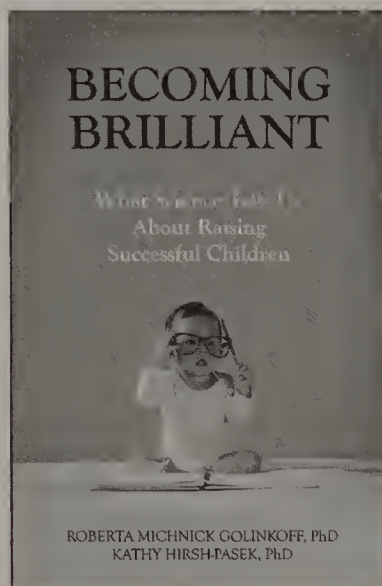


AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

# BECOMING BRILLIANT

What Science Tells Us About Raising Successful Children

Roberta Michnick Golinkoff, PhD, and Kathy Hirsh-Pasek, PhD



In just a few years, today's children and teens will forge careers that look nothing like those that were available to their parents or grandparents. While the U.S. economy becomes ever more information-driven, our system of education seems stuck on the idea that "content is king," neglecting other skills that 21st century citizens sorely need.

*Becoming Brilliant* offers solutions that parents can implement right now. Backed by the latest scientific

evidence and illustrated with examples of what's being done right in schools today, this book introduces the "6Cs"—collaboration, communication, content, critical thinking, creative innovation, and confidence—along with ways parents can nurture their children's development in each area. 2016. 344 pages. Paperback.

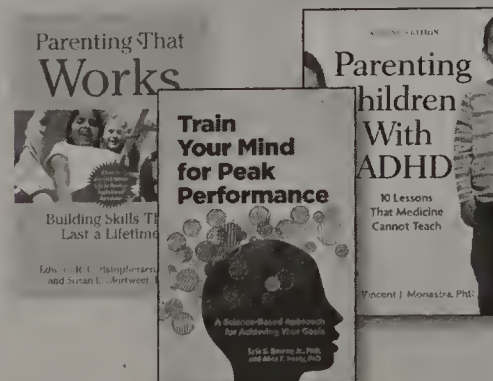
**An APA LifeTools® Book**

List: \$19.95 | APA Member/Affiliate: \$19.95 | ISBN 978-1-4338-2239-1 | Item # 4441027

## CONTENTS

Acknowledgments | Introduction | Chapter 1. Redefining Success in the 21st Century | Chapter 2. The Learning Industry and the Learning Sciences: How Educational Reform Sent Us in the Wrong Direction | Chapter 3. The Skills Needed for Success Are Global | Chapter 4. Hard Skills and Soft Skills: Finding the Perfect Balance | Chapter 5. Collaboration: No One Can Fiddle a Symphony | Chapter 6. Communication: Lines of Connection | Chapter 7. Toppling the King That Is Content | Chapter 8. Critical Thinking: What Counts as Evidence? | Chapter 9. Creative Innovation: Rearranging the Old to Make the New | Chapter 10. Confidence: Dare to Fail | Chapter 11. A Report Card for the 21st Century | Epilogue: What If? The Reprise | Notes | Index | About the Authors

## ALSO FROM APA LIFETOOLS®



### Parenting That Works

**Building Skills That Last a Lifetime**  
Edward R. Christophersen, PhD, ABPP,  
and Susan L. Mortweet, PhD

2003. 356 pages. Paperback.

List: \$16.95 | APA Member/Affiliate: \$16.95  
ISBN 978-1-55798-924-6 | Item # 4441003

AVAILABLE ON AMAZON KINDLE®

### Train Your Mind for Peak Performance A Science-Based Approach for Achieving Your Goals

Lyle E. Bourne, Jr., PhD,  
and Alice F. Healy, PhD

2014. 270 pages. Hardcover.

List: \$16.95 | APA Member/Affiliate: \$16.95  
ISBN 978-1-4338-1617-8 | Item # 4441021

AVAILABLE ON AMAZON KINDLE®

### Parenting Children With ADHD 10 Lessons That Medicine Cannot Teach

SECOND EDITION

Vincent J. Monastra, PhD

2014. 252 pages. Paperback.

List: \$16.95 | APA Member/Affiliate: \$16.95  
ISBN 978-1-4338-1571-3 | Item # 4441019

AVAILABLE ON AMAZON KINDLE®



PsychBOOKS®

Access to chapters from a variety  
of APA scholarly & professional books.

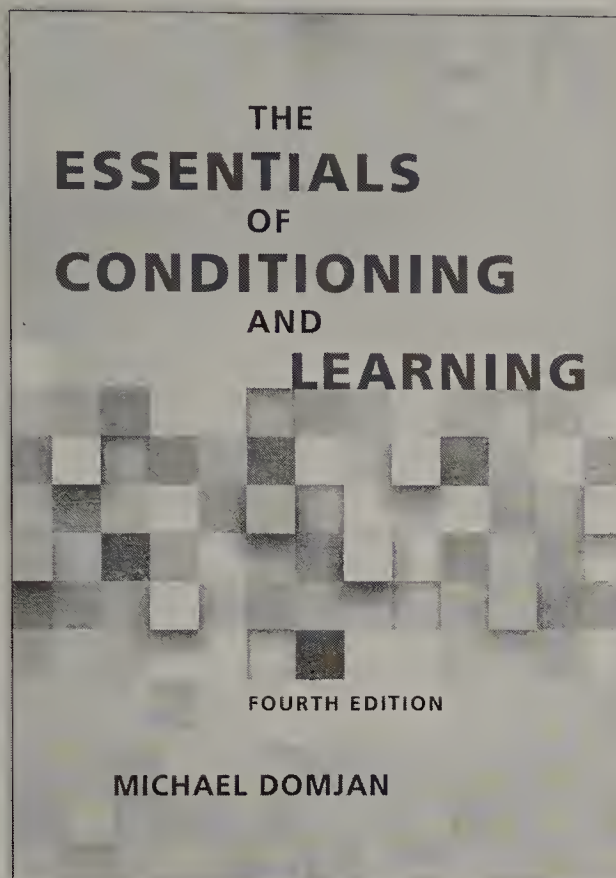
APA BOOKS ORDERING INFORMATION: 800-374-2721 • [www.apa.org/pubs/books](http://www.apa.org/pubs/books)

In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502

In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD3084





# The Essentials of Conditioning and Learning

FOURTH EDITION

*Michael Domjan*

Now in its fourth edition, Michael Domjan's classic textbook presents the basic principles of learning and conditioning in a concise and accessible style, with an emphasis on the latest influential research findings and theoretical perspectives. While the field of learning and conditioning is more than a hundred years old, new discoveries continue to be made and new applications of basic research are tackling major clinical problems. Domjan summarizes these developments as well as basic learning and conditioning principles using both human and animal examples. 2018. 376 pages. *Paperback*.

To view the full Table of Contents,  
visit [www.apa.org/h4313047](http://www.apa.org/h4313047)

List: \$64.95 | APA Member/Affiliate: \$54.95 | ISBN 978-1-4338-2778-5 | Item # 4313047

TO ORDER: 800-374-2721 ▪ [www.apa.org/h4313047](http://www.apa.org/h4313047)

In Washington, DC, call:

202-336-5510 ▪ TDD/TTY: 202-336-6123 ▪ Fax: 202-336-5502

In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972



An APA LifeTools® Book

## Raising Independent, Self-Confident Kids

Nine Essential Skills to Teach Your Child or Teen

Wendy L. Moss, PhD,  
and Donald A. Moses, MD

2018. 264 pages. Paperback.

List: \$19.95 | APA Member/Affiliate: \$19.95  
ISBN 978-1-4338-2825-6 | Item # 4441030

## Family Evaluation in Custody Litigation

Promoting Optimal Outcomes  
and Reducing Ethical Risks  
SECOND EDITION

G. Andrew H. Benjamin, Connie J. Beck,  
Morgan Shaw, and Robert Geffner

2018. 240 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-2831-7 | Item # 4317466

## Psychological Treatment of Cardiac Patients

Matthew M. Burg

2018. 160 pages. Paperback.

Series: *Clinical Health Psychology*

List: \$54.95 | APA Member/Affiliate: \$39.95  
ISBN 978-1-4338-2829-4 | Item # 4317469

## An ICD-10-CM Casebook and Workbook for Students

Psychological and Behavioral Conditions

Edited by Jack B. Schaffer  
and Emil Rodolfa

2018. 288 pages. Paperback.

List: \$34.95 | APA Member/Affiliate: \$34.95  
ISBN 978-1-4338-2827-0 | Item # 4311032

## Relational-Cultural Therapy

SECOND EDITION

Judith V. Jordan

2018. 190 pages. Paperback.

Series: *Theories of Psychotherapy Series®*

List: \$34.95 | APA Member/Affiliate: \$26.95  
ISBN 978-1-4338-2826-3 | Item # 4317467

## When Parents Are Incarcerated

Interdisciplinary Research and  
Interventions to Support Children

Edited by Christopher Wildeman,  
Anna R. Haskins,  
and Julie Poehlmann-Tynan

2018. 212 pages. Hardcover.

List: \$64.95 | APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-2821-8 | Item # 4318152

## Managing Your Research Data and Documentation

Kathy R. Berenson

2018. 112 pages. Paperback.

Series:

*Concise Guides to Conducting Behavioral,  
Health, and Social Science Research*

List: \$29.95 | APA Member/Affiliate: \$25.95  
ISBN 978-1-4338-2709-9 | Item # 4313048

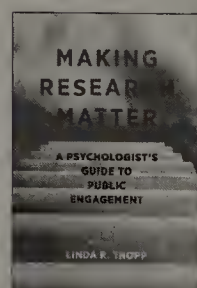
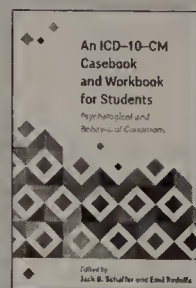
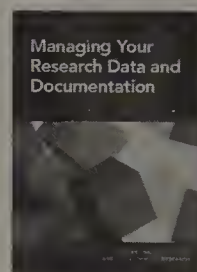
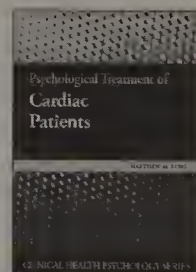
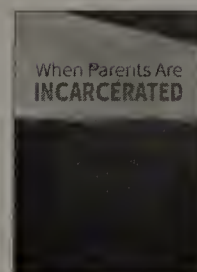
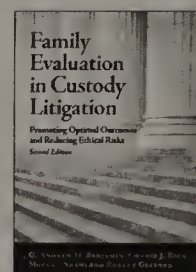
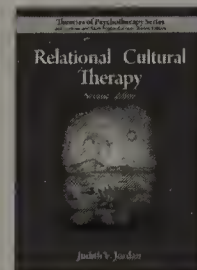
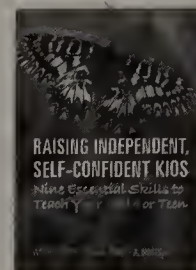
## Making Research Matter

A Psychologist's Guide  
to Public Engagement

Edited by Linda R. Tropp

2018. 208 pages. Paperback.

List: \$34.95 | APA Member/Affiliate: \$29.95  
ISBN 978-1-4338-2824-9 | Item # 4317468



► TO ORDER: 800-374-2721 | [www.apa.org/hn](http://www.apa.org/hn)



AMERICAN PSYCHOLOGICAL ASSOCIATION

AD3184